NeuroImage 263 (2022) 119598

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

# Neural correlates of multisensory enhancement in audiovisual narrative speech perception: A fMRI investigation

Lars A. Ross[a,b,c,*], Sophie Molholm[a,c], John S. Butler[c,d], Victor A. Del Bene[c,e], John J. Foxe[a,c,*]

[a] The Frederick J. and Marion A. Schindler Cognitive Neurophysiology Laboratory, The Ernest J. Del Monte Institute for Neuroscience, Department of Neuroscience, University of Rochester School of Medicine and Dentistry, Rochester, New York, 14642, USA
[b] Department of Imaging Sciences, University of Rochester Medical Center, University of Rochester School of Medicine and Dentistry, Rochester, New York, 14642, USA
[c] The Cognitive Neurophysiology Laboratory, Departments of Pediatrics and Neuroscience, Albert Einstein College of Medicine & Montefiore Medical Center, Bronx, New York, 10461, USA
[d] School of Mathematical Sciences, Technological University Dublin, Kevin Street Campus, Dublin, Ireland
[e] University of Alabama at Birmingham, Heersink School of Medicine, Department of Neurology, Birmingham, Alabama, 35233, USA

ARTICLE INFO

*Keywords:*
Subcortical
Naturalistic speech
Crossmodal
Semantic
Superior temporal gyrus
Neuroimaging
human

ABSTRACT

This fMRI study investigated the effect of seeing articulatory movements of a speaker while listening to a naturalistic narrative stimulus. It had the goal to identify regions of the language network showing multisensory enhancement under synchronous audiovisual conditions. We expected this enhancement to emerge in regions known to underlie the integration of auditory and visual information such as the posterior superior temporal gyrus as well as parts of the broader language network, including the semantic system. To this end we presented 53 participants with a continuous narration of a story in auditory alone, visual alone, and both synchronous and asynchronous audiovisual speech conditions while recording brain activity using BOLD fMRI. We found multisensory enhancement in an extensive network of regions underlying multisensory integration and parts of the semantic network as well as extralinguistic regions not usually associated with multisensory integration, namely the primary visual cortex and the bilateral amygdala. Analysis also revealed involvement of thalamic brain regions along the visual and auditory pathways more commonly associated with early sensory processing. We conclude that under natural listening conditions, multisensory enhancement not only involves sites of multisensory integration but many regions of the wider semantic network and includes regions associated with extralinguistic sensory, perceptual and cognitive processing.

## 1. Introduction

Sampling of information through multiple sensory systems enhances the likelihood of both detection and identification of survival-relevant objects or events in the environment. Inputs pertaining to the same objects or events are integrated across multiple stages of sensory and perceptual processing, leading to enhancements of behavior such as improved accuracy and faster reaction times for perceptual judgments (Bolognini et al., 2007; Brandwein et al., 2014; Brandwein et al., 2011; Diederich & Colonius, 2004; Foxe & Molholm, 2009; Frens et al., 1995; Molholm et al., 2004; Molholm et al., 2002; Nozawa et al., 1994; Rowland et al., 2007; Sperdin et al., 2009; Stein et al., 1989). Multisensory integration (MSI) organizes and reduces the complexity of our sensory environment by binding multiple sensory inputs into single, unified percepts and a failure of this function may lead to a sensory environment that is perceived as overwhelming with potential consequence of perceptual and behavioral deficits and maladaptive responses toward the environment (Ayres, 1979; Brandwein et al., 2015; Foxe & Molholm, 2009; Molholm et al., 2020).

One area of particular interest is speech recognition, where visual articulatory cues can strongly influence auditory speech perception (McGurk & MacDonald, 1976; Saint-Amour et al., 2007; Tjan et al., 2014) especially when the auditory speech signal is ambiguous, such as in noisy environments or in the presence of multiple simultaneous speakers (Benoit et al., 1994; Foxe et al., 2020; Foxe et al., 2015; Ma et al., 2009; MacLeod & Summerfield, 1987; Molholm et al., 2020; Richie & Kewley-Port, 2008; Ross et al., 2011; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007; Senkowski et al., 2008; Sumby, 1954). Despite the fact that

*List of abbreviations:* ATL, anterior temporal lobe; AV, audiovisual condition; AVa, asynchronous audiovisual condition; DMN, default mode network; dlPFC, dorsolateral prefrontal cortex; IFG, inferior frontal gyrus; MS, multisensory; MSI, Multisensory integration; MTG, middle temporal gyrus; pMTG, posterior medial temporal gyrus; STG, superior temporal gyrus; pSTS/G, posterior superior temporal sulcus/gyrus.

* Corresponding authors.
*E-mail addresses:* Lars_Ross@URMC.Rochester.edu (L.A. Ross), John_Foxe@URMC.Rochester.edu (J.J. Foxe).

https://doi.org/10.1016/j.neuroimage.2022.119598.
Received 18 February 2022; Received in revised form 26 August 2022; Accepted 28 August 2022
Available online 30 August 2022.

most of us are generally poor lip readers (Tye-Murray et al., 2007), the enhancing effects of visual speech can be dramatic, rendering mostly indecipherable vocalizations clearly audible (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007; Sumby, 1954). This well-known "principle of inverse effectiveness" (Meredith & Stein, 1986; Stein et al., 1988; Stein & Meredith, 1993) holds that multisensory enhancement generally increases with the degradation of the unisensory signals and has been shown across species (Stein et al., 1993) and experimental approaches (Sumby, 1954; van de Rijt et al., 2019) (James, 2012; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007; Stevenson et al., 2012). In the human brain the effect of congruent visual information can be observed at the neural level where low frequency neural activity phase locks to the temporal envelope of speech (Zion Golumbic et al., 2013) and has been shown to be enhanced in degraded auditory speech conditions (Crosse et al., 2016).

A common approach to investigating the neural mechanisms of AV speech processing and its enhancing effects is to use neuroimaging to compare hemodynamic responses to MS speech with responses to the constituent unisensory components (i.e., AV speech vs. auditory-alone or visual-alone speech). This allows for isolation of neural regions that show stronger responses to AV speech. The region most consistently localized is the superior temporal sulcus/gyrus (pSTS/G), an area well-known for its involvement in AV integration. This is the case whether the stimuli are as simple as nonsense monosyllables (Callan et al., 2003; Okada et al., 2013; Reale et al., 2007) and single words (Calvert et al., 1999; Wright et al., 2003), or as complex as a spoken story (Calvert et al., 2000). The MSI role of the pSTS/G extends to other aspects of AV speech stimuli such as congruency (Murase et al., 2008; Nath & Beauchamp, 2011), temporal synchrony (Macaluso et al., 2004; Noesselt et al., 2012), and ambiguity (Saint-Amour et al., 2007; Sekiyama et al., 2003; Stevenson & James, 2009). Other regions such as the primary auditory cortex (Calvert et al., 1999) and the motor cortex (Schomers & Pulvermuller, 2016) have also been implicated in AV speech processing suggesting more than one mechanism underlying observed MSI effects (Navarra, 2012). However, these areas have not been reliably implicated across studies and this lack of consistency regarding the regions that comprise the wider MSI speech network likely reflects, at least in part, the use of different paradigms, different stimulus materials and the use of different criteria to assess MS integration in the BOLD signal. However, we suspect that a major reason for between-study variability is that the studies often have relatively low levels of statistical power due to modest sample sizes. Notable exceptions are a large- scale lesion study ($N = 100$) on AV integration in speech (Hickok et al., 2018) and an fMRI study on functional connectivity between sensory and motor regions in audiovisual speech perception (Peelle et al., 2021).

The strong consensus regarding involvement of pSTS/G is based on highly stringent criteria. That is, this region survives many different types of experimental manipulations and statistical analysis methods and criteria across the many studies to examine assorted aspects of AV speech processing (Beauchamp, 2005; Calvert, 2001). However, speech is a complex stimulus and its processing engages a widely distributed network of regions serving a broad range of functions from sensory to semantic processing (Hickok & Poeppel, 2007; Price, 2010; Rauschecker, 2012). As such, the visual benefit manifested in enhanced speech perception is unlikely to be related solely to the involvement of a single region but rather to coordinated activity across the network of speech processing regions including perisylvian language areas and the motor and premotor cortex. It has been shown that MS interactions occur at multiple stages of information processing (Foxe & Schroeder, 2005) and a number of studies have reported that AV speech amplifies activity in early primary auditory cortex (Callan et al., 2003; Calvert et al., 1999; Calvert et al., 1997; Calvert et al., 2000; Okada et al., 2013), visual motion regions (Puce et al., 1998; Puce et al., 2003; Wright et al., 2003; Yarkoni et al., 2011), and prefrontal regions such as Broca's area and premotor cortex (Iacoboni, 2008; Meister et al., 2007; Ojanen et al., 2005; Skipper et al., 2005; Wilson et al., 2004).

Further, it is reasonable to assume that under natural listening conditions, integration in MS regions has downstream consequences in the larger speech and language network. Moreover, most studies investigating AV integration used truncated speech material such as syllables and words often in the context of a McGurk-type paradigm where participants are asked to reconcile conflicting auditory and visual cues. It has been questioned whether these tasks engage the same mechanisms active in more natural AV speech processing (Alsius et al., 2018; Hickok et al., 2018; Peelle, 2019; Van Engen & Peelle, 2014). This case has also been made in regard to the investigation of cortical entrainment as a central mechanism for speech perception where low- frequency oscillatory activity aligns with the envelope of the acoustic speech stimulus (Alexandrou et al., 2020; Lakatos et al., 2019; Lakatos et al., 2008; Schroeder & Lakatos, 2009) (Ding & Simon, 2014; Haegens & Golumbic, 2018; Zoefel et al., 2018).

We therefore expect these broader MS enhancement effects to arise with natural and more complex stimulus material such as narratives (Hamilton & Huth, 2020; Hasson et al., 2018; Huth et al., 2016). For the purpose of this investigation, we use the term MS enhancement in a broader sense to refer to processes of MS integration and their possible consequences on linguistic and cognitive processing because our experimental approach does not strictly distinguish them from one another.

Therefore, the present study had several central goals. The first was to comprehensively map the network of brain regions involved in audiovisual enhancement in natural narrative speech perception in a large sample of healthy adults by comparing brain responses to an audiovisual speech stimulus to the responses to the constituent unisensory responses presented in isolation. We also employed an additional approach that has been used in the past in behavioral, hemodynamic and electrophysiolocal studies to study MS integration by comparing audiovisual aligned with misaligned audiovisual stimulus conditions (Miller & D'Esposito, 2005) (Stevenson et al., 2010; van Atteveldt et al., 2007; van Wassenhove et al., 2007). Importantly, this study focused not only on identifying regions of AV integration but also aimed to assess the downstream effects of AV integration on the larger language network.

Using BOLD fMRI, we presented continuous natural speech in varying conditions: auditory alone (A), visual alone (V), synchronous audiovisual (AV) and an asynchronous audiovisual condition (AVa). We characterized regions engaged in the presentation of the unisensory conditions (A and V) in order to identify brain regions engaged in natural narrative speech processing and speechreading respectively. We mapped AV enhancement effects by examining areas that responded more strongly to the AV speech compared to unisensory speech stimuli. We employed the maximum criterion (Beauchamp, 2005; James, 2012) by performing a conjunction analysis identifying regions in which the AV- response was significantly larger than the A and the V response [(AV > A) ∧ (AV > V)] while constraining the analysis to regions in which AV was significantly larger than baseline (AV > 0). We expected to observe enhancement in regions known to be involved in MS integration (Erickson et al., 2014) (Calvert & Thesen, 2004) and in regions downstream from known MSI sites reflecting the effects of successful integration in the larger speech processing network. We also explored which regions of the identified network showed a superadditive response to the AV stimulus [AV > (A + V)]. This criterion is considered to be much more conservative than the maximum criterion (Beauchamp, 2005; James, 2012) and we therefore expected enhancement to be constrained to "classic" MS integration sites such as the pSTS. We also added an experimental condition where the AV inputs were out of synchrony and compared responses to synchronous and asynchronous AV speech. The purpose was to investigate regions sensitive to the temporal alignment of the AV speech signals, under the assumption that MS binding occurs when sensory signals are correlated in time (Stein et al., 1988). Based on previous literature (Stevenson et al., 2010), we expected effects of synchrony and asynchrony to emerge within the medial STS.

In a final exploratory analysis, we assessed the relationship between activation to the respective conditions in the fMRI experiment and behavioral measures on an AV speech perception task obtained from the same subjects in an experiment completed outside the scanner. The goal was to test whether hemodynamic correlates of AV enhancement were related to the ability to benefit from visual articulation in an audiovisual speech perception task.

## 2. Materials and methods

### 2.1. Participants

From an original sample of 60 participants, 7 were excluded based on technical difficulties during the scan or post processing, due to excess motion or lack of task compliance. The data of 53 native English-speaking adults with no history of neurological and psychiatric problems and no substance abuse (25 female, age range = 20 -to 35 years, $M = 25$ years, $SD = 3.8$ years) were included in the following fMRI analyses. All had normal hearing and normal or corrected-to-normal vision. Out of the 53, 47 were right-handed, 3 left-handed and 2 were ambidextrous (Oldfield, 1971). Handedness of one participant was not recorded. The study was approved by the Institutional Review Board of the Albert Einstein College of Medicine and all procedures were conducted in accordance with the tenets of the Declaration of Helsinki. All participants gave written informed consent and were paid for their participation.

### 2.2. MRI acquisition

Imaging data were acquired using a 3.0 Tesla Philips Achieva TX scanner with a 32-channel head coil. A T1-weighted whole-head anatomical volume was obtained using a 3D magnetization-prepared rapid gradient-echo (MP-RAGE) sequence (echo time [TE] = 3.7 ms, repetition time [TR] = 8.2 ms, flip angle [FA] = 8 degrees, voxel size = 1 × 1 × 1 $mm^3$, matrix = 256 × 256, FOV = 256 × 256 $mm^2$, number of slices = 220). $T2^*$-weighted functional scans were acquired using gradient echo-planar imaging (EPI). This acquisition covered the whole brain excluding inferior aspects of the cerebellum below the horizontal fissure (axial acquisition in ascending order, TE = 20 ms, TR = 2000 ms, FA = 90 degrees, voxel size = 1.67 × 1.67 × 2.30 $mm^3$, matrix = 144 × 144, FOV = 240 × 240 $mm^2$, number of slices per volume = 50, total number of volumes = 158 (run1) + 172 (run2) + 146 (run 3).

### 2.3. fMRI task

Participants were presented with video recordings of a speaker reading from a children's story about economic and environmental issues called "The Lorax" written by Dr. Seuss. The story was narrated by an adult female, caucasian actor speaking directly into the camera (0 degree angle) as if directly speaking to a listener with continuous eye contact. The video was recorded in a quiet, well lit room with the actor standing before a plain grey background at the center of the screen with only her head and torso visible (see Figure 4 in the appendix). The video of the story (lasting 14 min 38 s) was segmented into sections of varying length ranging from 8 to 22 s. The length of the blocks was determined by natural break points in the narration to ascertain smooth transitions between blocks while considering block length as a factor in fMRI design efficiency (Maus et al., 2010; Smith et al., 2007). The frame rate of the video recordings was 29 frames per second. Each section was randomly assigned for each participant to one of four conditions: auditory (A), visual (V), synchronous audiovisual (AV), and asynchronous audiovisual (AVa). As such, block length is a random variable that is not associated with a given condition. The A and V conditions presented the auditory and visual stimuli alone, respectively. During the A condition, an unedited still image of the speaker looking directly into the camera with a neutral facial expression was presented and participants were told to look at the picture while listening to the story.

The AV and AVa conditions presented both the auditory and visual stimuli, but in the AV condition, the two inputs were presented in synchrony whereas in the AVa conditions, the visual input was delayed by 400 ms relative to the timing of the auditory input such that the audio and video were clearly misaligned. The full story was presented in 3 runs of 4 min 50 s, 5 min 20 s, and 4 min 28 s, respectively. Participants were instructed to follow the whole story carefully regardless of the changing presentation mode. The story in each run was followed by a resting period during which a screen containing a sign saying "please relax" was presented briefly and disappeared, leaving only a blank screen. Participants were asked to rest during this period with their eyes open. The resting period lasted 18, 16 and 16 s for the respective 3 runs without rest periods between blocks. For a given contrast the baseline therefore represents the average time course. Retention of the story content was assessed with a 10-item, four- option multiple choice questionnaire after the scan which can be found in the appendix. Note that this experiment was designed with an eye towards future investigations of MSI processes across development and was therefore constructed to be suitable for use in children (hence the choice of a narrative that would appeal to all age groups). The presentation of a continuous narrative precluded the use of a simultaneous behavioral task, so our intention here was to ensure task compliance via our instruction that the subject would be "tested" after the scan. We included the five adults for whom we did not have the questionnaire data because 1), eye-tracking measures in these individuals made it clear that they fixated the screen appropriately with eyes open throughout the experiment, and 2), we inspected the statistical maps for each subject to ensure the presence of typical auditory and visual sensory activation patterns indicating compliance with experimenter instructions.

Throughout the whole MRI session, participants wore foam ear plugs to attenuate the scanner noise and MR-compatible headphones (the Serene Sound system; Resonance Technology, Inc.) through which the auditory stimuli were presented (bit rate: 1536 kbps; sample rate: 48000 kHz). The SPL of the headphones was kept constant in the range of 90 to 95 dB across the participants who reported this volume to be audible and comfortable. Participants wore MR-compatible glasses (the VisuaStim Digital system; Resonance Technology, Inc.) through which the visual stimuli were delivered at a refresh rate of 60 Hz. An eye tracker (the MReyetracking system; Resonance Technology, Inc.) was mounted inside the glasses and used to monitor that participants' eyes were open and watching the video, throughout the task.

### 2.4. fMRI analysis

All imaging data were analyzed in BrainVoyager (version 22.2, Brain Innovation, Maastricht, the Netherlands). The functional data were preprocessed using interscan slice time correction (cubic spline interpolation) and 3D rigid-body motion correction (trilinear sinc interpolation). The data of all three runs were aligned to the first volume of the first run. No subject data were removed for excess motion based on a cutoff of 2mm/degrees in any direction). Individual anatomical images were transformed into Talairach space (sinc interpolation) and functional imaging data were aligned to the individual's anatomy using boundary- based registration (Greve & Fischl, 2009) and inspected for quality of registration. The time courses for each participant were subsequently temporal high pass filtered with a GLM Fourier basis set and spatially smoothed using a 6mm FWMH Gaussian Kernel before transformation into Talairach space.

Voxel-wise statistical analyses were performed on the (%) normalized functional data using a two-level random-effects GLM approach with A, V, AV and AVa as predictors which were convolved with a standard two-gamma hemodynamic response function. We used a Talairach mask to exclude voxels outside the brain.

The following contrasts of interest and conjunction analyses were performed and reported: (1) A vs baseline: This analysis was performed to identify brain regions active during the processing of the story (narrative) without visual articulatory information. (2) V vs baseline: Here, we investigated regions involved in the processing of visual articulatory information. (3) [(AV-A) ∧ (AV-V)]: This was the analysis critical for the identification of MS enhancement according to the Max Criterion (Beauchamp, 2005; James, 2012) and tests via mathematical conjunction of the (AV-A) and (AV-V) contrasts (Nichols et al., 2005) whether activation to the AV condition significantly supersedes the A and the V condition against their baseline. In regions meeting this criterion activation to the AV condition is significantly larger than activation to the A condition *and* activation to the V condition (4) (AV > A+V): We also tested AV enhancement according to the additive (superadditive) criterion (Calvert & Thesen, 2004) where the BOLD response to the AV condition was larger than the sum of the A and V responses. For this analysis we summed the normalized predictor values for the A and V conditions from each voxel and subtracted them from the predictor values of the AV condition. The resulting values were tested against zero using a t-test. (5) AV vs. AVa: In this contrast we compared the synchronous AV and the asynchronous AV condition.

The following analyses were secondary in regard to the goals of this study and are reported in the appendix: (6) A vs V: The difference between auditory and visual conditions. This contrast is particularly sensitive to activations in the auditory and visual cortices and was applied on a single subject basis after a fixed effects GLM with the predictors of interest to assure the compliance to the experimental instructions and the absence of failed data acquisition due to technical problems. On a group level, this analysis was performed to delineate regions where both conditions differed from one another and allow a comparison to the statistical map of regions where they were active in conjunction, as follows. (7) A ∧ V: The conjunction of the contrasts (Nichols et al., 2005) of A and V conditions against baseline tests for voxels in which both A and V conditions differ significantly from baseline. We were interested in this analysis primarily to determine whether regions of MS enhancement are also responsive to the A and V conditions.

For all whole brain analyses we used the false discovery rate (FDR) procedure (Genovese et al., 2002) to control for multiple comparisons at $q < 0.05$.

### 2.5. Out of scanner MS speech recognition behavioral task

Stimulus materials consisted of digital recordings of 300 simple monosyllabic words spoken by a female speaker. This set of words was a subset of the stimulus material created for a previous experiment in our laboratory (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007) and used in a previous study (Ross et al., 2011). These words were taken from the "MRC Psycholinguistic Database" (Coltheart, 1981) and were selected from a well-characterized normed set based on their written-word frequency (Kucera & Francis, 1967). The subset of words for the present experiment is a selection of simple, high-frequency words likely to be in the lexicon of participants in the age-range of our sample. The recorded movies were digitally re-mastered so that the length of the movie (1.3 sec) and the onset of the acoustic signal were similar across all words. Average voice onset occurred at 520ms after movie onset (SD= 30ms). The words were presented at approximately 50dBA FSPL, at seven levels of intelligibility including a condition with no noise (NN) and six conditions with added pink noise at 50, 53, 56, 59, 62 and 65dBA FSPL sound pressure. Noise onset was synchronized with movie onset. The signal-to-noise ratios (SNRs) were therefore NN, 0, -3, -6, -9, -12, –15dBA FSPL. These SNRs were chosen to cover a performance range in the auditory-alone condition from 0% recognized words at the lowest SNR to almost perfect recognition performance with no noise. The movies were presented on a monitor (NEC Multisync FE 2111SB) at 80cm distance from the eyes of the participants. The face of the speaker extended approximately 6.44° of visual angle horizontally and 8.58° vertically (hairline to chin). The speaker looked straight (no angle) at the camera with a neutral facial expression. A still image of one of the videos is shown in Supplementary Figure 5 in the appendix. The words and pink noise were presented over headphones (Sennheiser, model HD 555).

The main experiment consisted of three randomly intermixed conditions: In the auditory-alone condition (A-alone) the auditory words were presented in conjunction with a still image of the speakers' face; in the AV condition the auditory words were presented in conjunction with the corresponding video of the speaker articulating the words. Finally, in the visual alone condition (V-alone) only the video of the speaker's articulations was presented. The word stimuli were presented in a fixed order and the condition (the noise level and whether it was presented as A-alone, V-alone or AV) was assigned to each word randomly. Stimuli were presented in 15 blocks of 20 words with a total of 300 stimulus presentations. There were 140 stimuli for the A and AV conditions respectively (20 stimuli per condition and intelligibility level) and 20 stimuli for the V condition that was presented without noise.

Task: Participants were instructed to watch the screen and verbally report which word they heard (or saw in the V-alone condition). If a word was not clearly understood, participants were encouraged to make their best guess. An experimenter, seated approximately 1 m distance from the participant at a 90° angle to the participant-screen axis, monitored participant's adherence to maintaining fixation on the screen. The experimenter recorded the participants' responses which were later scored for correctness. Only responses that exactly matched the presented word were considered correct. Any other response was recorded as incorrect.

#### 2.5.1. Analyses of task performance

We submitted percent correct responses for each condition to a repeated measures analysis of variance (RM-ANOVA) with factors of stimulus condition (A vs. AV), SNR level (7 levels) and biological sex as a between subjects' factor as well as age as a covariate. Performance in the V-alone condition was analyzed separately because it was only presented without noise. Violations of the sphericity assumption of the RM-ANOVA were corrected by adjusting the degrees of freedom with the Greenhouse-Geisser correction method. MS enhancement (or AV-gain) was operationalized here as the difference in performance between the AV and the A-alone condition (AV – A-alone). This analysis was performed at the four lowest SNRs because the variance at higher SNRs becomes increasingly constrained by ceiling performance (Ross et al., 2011). We performed two- tailed Pearson correlation tests between the A and the AV conditions at the four lowest SNRs (average) to determine if A- performance under noisy conditions was negatively associated with AV- performance at the same SNRs. We also tested for an association between the A and V conditions.

Finally, we also tested the hypothesis that individuals with more difficulty perceiving auditory speech when it is masked with noise are better speech-readers, who therefore benefit more from AV input. The presence of this trade-off gained recent support from a study showing that early electrophysiological indices of auditory processing predict auditory, visual and AV speech processing (Dias et al., 2021b). If such effects were apparent in our data, our goal was to investigate possible relationships of auditory processing ability with measures of brain activity in our fMRI study.

#### 2.5.2. Correlation with BOLD measures

For the analysis of MS enhancement we averaged MS gain at the four lowest SNRs where most audiovisual enhancement was observed and computed voxel-wise correlations with the beta weights of the AV-A contrast of our BOLD data. The resulting correlation (Pearson's r) maps were thresholded at $p = 0.001$ as suggested by recent findings (Eklund et al., 2016) in order to control for family wise error rate (FWE). If this initial threshold produced a map lacking a sufficient size and distribution of significant clusters, this threshold was iteratively increased to a maximum of $p = 0.01$. This compromise in regard to the risk for

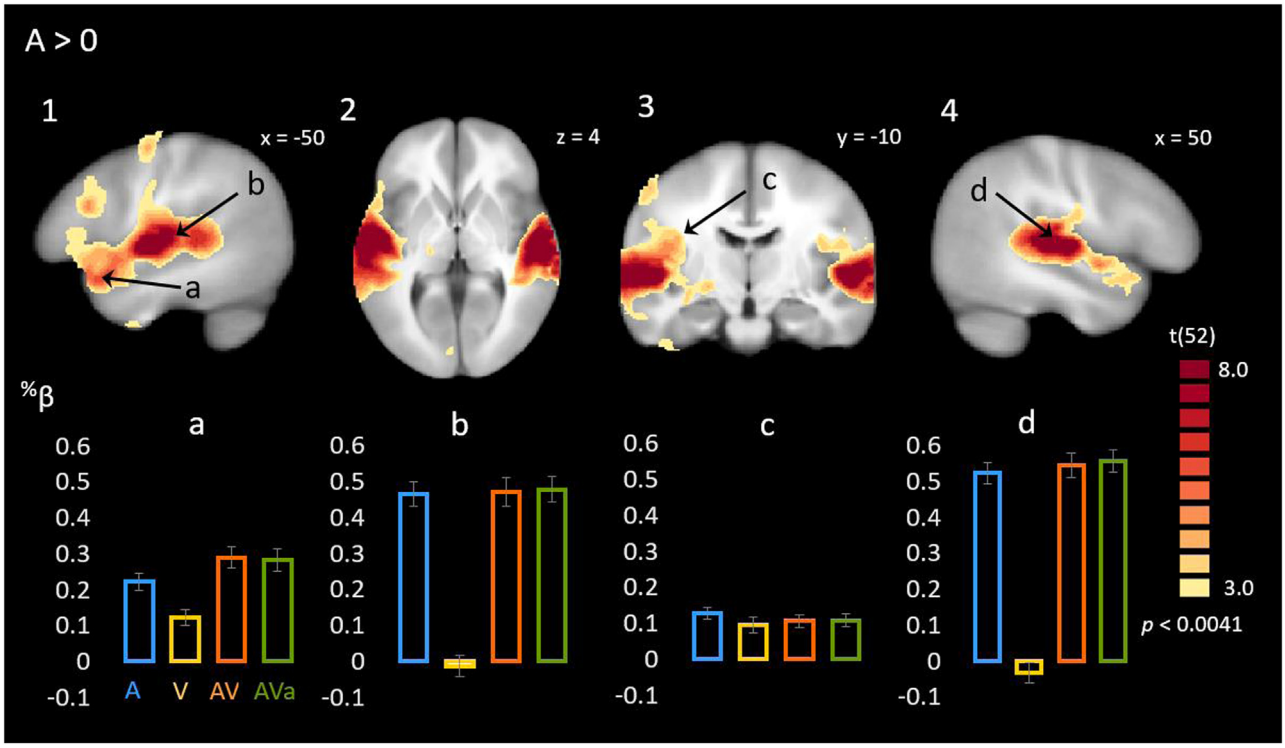


**Fig. 1.** Statistical comparison of the A condition to baseline.
Maps 1-4 show voxels with significant t-scores of the comparison of the A- condition to baseline FDR- corrected (q = 0.05) for multiple comparisons. Bar graphs represent selected % transformed predictor values for A, V, AV and AVa conditions averaged over 4 functional voxels centered around peak voxel locations (see Table 1). a) Left anterior superior temporal gyrus; b) left Heschl's gyrus; c) left insula; d) right Heschl's gyrus.

false positives was motivated by our expectation that the effect size of the correlation of our GLM predictors with measures of behavioral performance would not be of the same magnitude as the effect size of moderate BOLD effects for which a $p < 0.001$ threshold was shown to be appropriate. We therefore did not expect this threshold would result in a statistical map with a cluster distribution suitable for a subsequent Monte Carlo simulation. We used the thresholded map as input for the Cluster- Level Statistical Threshold Estimator plugin in Brainvoyager using 5000 iterations. This tool simulates the distribution of normally distributed noise based on the smoothness of the map used as input in each iteration step and records the frequency and size of the resulting clusters. We performed exploratory analyses of the relationship between BOLD effects and behavioral performance in the A, V and AV conditions.

Finally, we explored relationships between questionnaire performance and BOLD measures. This analysis is secondary to the aims of this study and is reported in the appendix.

## 3. Results

### 3.1. Auditory alone (A)

#### 3.1.1. Major findings

The stimulation in the auditory alone condition (A vs. baseline) resulted in strong activation in bilateral Heschl's gyrus with peak activations within the primary auditory cortex (see Fig. 1 and Table 1). From these locations, clusters in both hemispheres extended anteriorly along the superior temporal gyrus and its upper bank into the anterior temporal lobes (ATLs, Fig. 1, panels 1 and 4) including a cluster in the ventral ATL in the left hemisphere (Fig. 1, panels 1 and 3), laterally along the transverse temporal gyrus (Fig. 1, panel 3) and posteriorly toward the posterior parietal junction. Activations extended from the primary auditory cortices into the ventral motor cortex along the roofs of the lateral sulci covering the parietal operculae in both hemispheres (Fig. 1, panel 3). The auditory condition also engaged left hemispheric regions in the frontal lobe including the inferior frontal gyrus (IFG, Fig. 1, panel 1), the dorsolateral prefrontal cortex (dlPFC, Fig. 1, panel 1). The supplementary motor cortex was engaged in both hemispheres (Fig. 1, panel 1, right hemispheric cluster not shown).

#### 3.1.2. Minor findings

Also, in the left hemisphere we found a cluster in lentiform nucleus with a center in the globus pallidus extending laterally into the putamen and nearby in the ventral posterior lateral nucleus of the thalamus (Fig. 1, panel 2 and 3). Smaller clusters were found in the bilateral middle temporal gyrus and the left lingual gyrus (Fig. 1, panel 2) as well as the left cerebellar hemisphere and the vermis (not shown). We also found clusters in cerebral white matter in the genu and splenium of the corpus callosum at the borders of the anterior and posterior horn of the left ventricle and the body of the corpus callosum at the midline (not shown). We also found a small cluster of activation in the right crus cerebri of the cerebellar peduncles (not shown). Upon close inspection these white matter clusters did not appear to be the result of "spill over" from nearby grey matter regions. Finally, BOLD activity in the primary visual cortex in this condition was significantly below baseline.

### 3.2. Visual alone (V)

In line with our expectations, the visual alone stimulation resulted in a strong BOLD response in primary visual cortices of bilateral occipital poles (Fig. 2, panels 2 and 3, Table 2). The clusters extended laterally to form two prominent foci of activation in the lateral occipital cortices (LOC) (Fig. 2, panels 1, 2 and 5). From the left LOC region, significant BOLD activity appeared to follow along the ventral visual pathway into the fusiform gyrus (Fig. 2, panel 4) and dorsal visual pathway toward the

**Table 1**
Clusters of significant activity resulting from the contrast of the A condition vs. baseline.
Significant clusters are numbered and reported with their t-statistic and location in Talairach space in the order of cluster size. In cases where clusters spanned over more than one anatomical or functional region additional peak voxels are reported together with their corresponding anatomical region.

| **A > 0** Cluster | | L/R | *t*-statistic | x | y | z | Voxels |
|---|---|---|---|---|---|---|---|
| 1 | Heschl's gyrus | L | 12.83 | -48 | -13 | 3 | 7317 |
| | Superior temporal gyrus (posterior division) | L | 11.81 | -64 | -16 | 9 | |
| | Superior temporal gyrus (anterior division) | L | 6.68 | -48 | 12 | -13 | |
| | Insula /Operculum | L | 11.2 | -42 | -6 | 18 | |
| | Lentiform nucleus (putamen) | L | 4.12 | -31 | -9 | -5 | |
| | Lentiform nucleus (lateral globus pallidus) | L | 4.55 | -21 | -13 | -2 | |
| | Thalamus (vent.post.lat. nucleus) | L | 3.82 | -17 | -18 | 4 | |
| 2 | Heschl's gyrus | R | 13.52 | 54 | -9 | 3 | 4046 |
| | Superior temporal gyrus (posterior division) | R | 12.54 | 64 | -19 | 6 | |
| | Superior temporal gyrus (anterior division) | R | 4.32 | 47 | 17 | -9 | |
| | Insula | R | 4.16 | 39 | -12 | 20 | |
| 3 | Splenium of the corpus callosum | L | 4.70 | -19 | -45 | 15 | 194 |
| 4 | Anterior inferior temporal gyrus | L | 4.21 | -36 | -8 | -39 | 165 |
| 5 | Precentral gyrus | R | 4.57 | 57 | -6 | 44 | 32 |
| 6 | Genu of the corpus callosum | L | 3.80 | -17 | 25 | 17 | 28 |
| 7 | Crus cerebri cerebellar peduncles | R | 4.07 | 14 | -15 | -15 | 20 |
| 8 | MTG | L | 3.67 | -62 | -44 | -12 | 20 |
| 9 | IFG | L | 3.38 | -44 | 6 | 23 | 17 |
| 10 | Body of the corpus callosum | R | 3.47 | 2 | 3 | 21 | 15 |

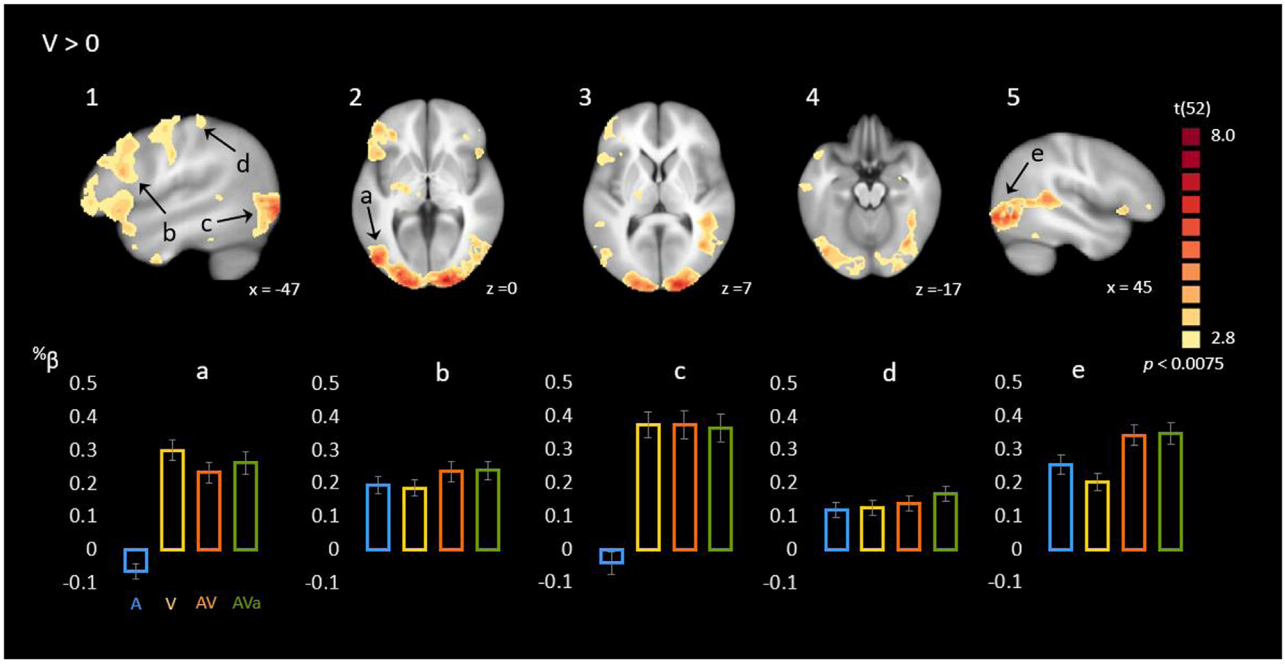


**Fig. 2.** Statistical comparison of the V condition to baseline.
Maps 1-5 show voxels with significant t-scores of the comparison of the V- condition to baseline FDR- corrected (q = 0.05) for multiple comparisons. Bar graphs represent selected % transformed predictor values for A, V, AV and AVa conditions averaged over 4 functional voxels centered around peak voxel locations (see Table 2). a) Left LOC; b) left IFG; c) left occipital pole; d) left precentral gyrus; e) right pSTS.

occipito-temporal cortex passing visual motion area MT and MST and extending into the pSTS/G (Fig. 2, panel 5). In the left hemisphere significant BOLD effects of the ventral pathway did not extend anteriorly as far as in the right hemisphere and spared the posterior temporal cortex but showed a small cluster in the pSTS/G (Fig. 2, panel 1). Ventral aspects of the ATLs in both hemispheres showed significant responses but only the left hemisphere also showed clusters in the middle and superior ATL (Fig. 2, panels 1 and 4).

There was widespread significant activation in the left frontal lobe (Fig. 2, panels 1, 2 and 3) involving the ventrolateral prefrontal and lateral frontopolar cortex along the IFG, the nearby frontal operculum and the medial frontal gyrus near the midline. We also found significant clusters in the dlPFC, ventral premotor and ventral and dorsal motor regions. The activations in the right frontal lobe (Fig. 2, panels 2 and 5) were smaller than in the left hemisphere and included the IFG and lateral frontopolar cortex and primary motor regions. Like in the A>0 contrast, we found a cluster in the left lentiform nucleus and thalamus (Fig. 2, panels 2 and 3).

Finally, we found significant activity in the right amygdala (Fig. 2, panel 4) and the cerebellar vermis (not shown). As in the A > 0 contrast we found activity in the white matter of the splenium of the corpus callosum at the border to the left lateral ventricle (not shown).

**Table 2**
Clusters of significant activity resulting from the contrast of the V condition vs. baseline.
Significant clusters are numbered and reported with their t-statistic and location in Talairach space in the order of cluster size. In cases where clusters spanned over more than one anatomical or functional region additional peak voxels are reported together with their corresponding anatomical region.

| **V > 0** Cluster | | L/R | *t*-statistic | x | y | z | Voxels |
|---|---|---|---|---|---|---|---|
| 1 | Occipital pole/cuneus | L | 9.12 | -20 | -95 | 2 | 8548 |
| | Occipital pole/cuneus | R | 9.27 | 17 | -95 | 0 | |
| | Lateral occipital complex | L | 8.98 | -43 | -75 | -5 | |
| | Lateral occipital complex | R | 6.76 | 45 | -63 | -5 | |
| | pSTS | R | 5.47 | 42 | -37 | 8 | |
| 2 | IFG | L | 5.96 | -58 | 15 | 22 | 3036 |
| | Precentral gyrus | L | 4.72 | -53 | -5 | 43 | |
| 3 | Lentiform nucleus (lateral globus pallidus) | L | 4.27 | -21 | -12 | -2 | 265 |
| 4 | MTG | L | 4.36 | -63 | -15 | -13 | 97 |
| 5 | Ventral ATL | R | 4.34 | 37 | -9 | -33 | 89 |
| 6 | Medial frontal gyrus | L | 3.83 | -9 | 33 | 30 | 76 |
| 7 | Amygdala | R | 3.41 | 28 | -1 | -12 | 51 |
| 8 | Pre- and Postcentral gyrus | L | 3.77 | -48 | -29 | 51 | 49 |
| 9 | White matter lateral ventricle | L | 3.45 | -19 | -45 | 16 | 46 |
| 10 | IFG | R | 4.8 | 47 | 21 | -1 | 43 |
| 11 | Medial frontal gyrus | L | 4.24 | -15 | 41 | 45 | 41 |
| 12 | ventral ATL | L | 4.02 | -50 | -3 | -35 | 41 |
| 13 | pSTS/G | L | 3.30 | -51 | -39 | 7 | 38 |
| 14 | Precentral gyrus | R | 4.02 | 57 | -6 | 42 | 35 |
| 15 | Parahippocampus | L | 3.81 | -42 | -29 | -13 | 35 |
| 16 | Culmen | L | 3.49 | -4 | -45 | -9 | 28 |
| 17 | IFG | R | 3.31 | 46 | 39 | 3 | 21 |
| 18 | Thalamus | L | 3.31 | -8 | -18 | 1 | 19 |
| 19 | IFG | R | 3.53 | 34 | 34 | 0 | 16 |
| 20 | Anterior MTG | L | 3.41 | -47 | 13 | -24 | 16 |
| 21 | Precentral gyrus | R | 3.34 | 50 | -9 | 30 | 15 |

**Table 3**
Clusters of significant activity (Max. criterion) resulting from the conjunction between the AV-A and AV-V contrasts. Significant clusters are numbered and reported with their t-statistic and location in Talairach space in the order of cluster size. In cases where clusters spanned over more than one anatomical or functional region additional peak voxels are reported together with their corresponding anatomical region.

| **(AV-A) ^ (AV-V)** Cluster | | L/R | *t*-statistic | x | y | z | Voxels |
|---|---|---|---|---|---|---|---|
| 1 | Thalamus (LGN) | R | 6.57 | 21 | -24 | 0 | 3231 |
| | Thalamus (MGN) | R | 5.936 | 9 | -27 | 1 | |
| | Amygdala | R | 4.88 | 21 | -5 | -11 | |
| | Thalamus (LGN) | L | 4.64 | -22 | -23 | -2 | |
| | Thalamus (MGN) | L | 3.24 | -9 | -26 | -1 | |
| | Amygdala | L | 5.34 | -20 | -5 | -13 | |
| 2 | pSTS | R | 6.74 | 44 | -37 | 8 | 2300 |
| | Anterior STS | R | 5.81 | 46 | 12 | -18 | |
| 3 | pSTS | L | 6.31 | -49 | -39 | 6 | 1158 |
| | Anterior STS | L | 4.81 | -49 | 15 | -18 | |
| 4 | Cuneus/ occipital pole | L | 4.36 | -5 | -93 | 7 | 566 |
| 5 | Precentral gyrus | R | 3.41 | 53 | -2 | 46 | 28 |
| 6 | IFG | L | 3.36 | -60 | 19 | 20 | 23 |
| 7 | Lingual gyrus | R | 3.33 | 16 | -84 | -1 | 22 |
| 8 | IFG | R | 3.34 | 51 | 22 | 16 | 17 |

### 3.3. MS enhancement: Max. criterion [(AV-A) ∧ (AV-V)]

The purpose of this conjunction analysis was to identify regions showing MS enhancement where the BOLD response to the AV condition was greater than to the auditory and visual condition respectively (Max criterion). We limited this analysis to regions where AV was greater than baseline (Fig. 3, Table 3).

We found large MS activations along the STS in both hemispheres spanning from the ATLs into the posterior STS (Fig. 3, panels 1, 4, 6). The posterior sections of these two large clusters extend dorsally to cover the posterior STG and the supramarginal gyrus. While these activations are represented by continuous clusters, they are likely to represent functionally distinct regions. We therefore increased the statistical threshold in a stepwise fashion from the FDR corrected threshold at $p = 0.0067$ to $p = 0.000002$ to identify local peak activations within the larger STS clusters (not shown in Figure). We found that both STS clusters contained an anterior, medial and posterior peak in both hemispheres and within the supramarginal gyrus in the left hemisphere.

We found significant MS gain in ventral parts of the left temporal lobe (not shown) and particularly in the bilateral amygdalae (Fig. 3, panel 6). Further, and surprising to us, were bilateral twin clusters in the posterior thalamus encompassing the medial geniculate nuclei, lateral geniculate nuclei and the pulvinar (Fig. 3, panel 5). The conjunction was also significant in several regions of the frontal cortex (not shown). These included two smaller clusters in the bilateral IFG and one in the right precentral gyrus (premotor cortex). Finally, the statistical conjunction of audiovisual enhancements over unisensory activations also revealed a significant engagement of the bilateral occipital poles (Fig. 3, panel 4).

### 3.4. MS enhancement: Superadditivity AV > (A+V)

Here, we examined the distribution of superadditivity in regions where the AV condition was above baseline (Fig. 4, Table 4). The FDR-corrected map shows significant superadditivity within the bilateral STS encompassing primary and secondary auditory cortices and more anterior in the STS reaching into the ATL in the right hemisphere (Fig. 4, panels 1 and 4). The extracted % transformed beta weights show that in the primary auditory cortices, the AV condition does not significantly exceed the A-condition in the left hemisphere (Fig. 4, panel 1, bar graph e) and in the right hemisphere (Fig. 4, panel 4, bar graph c). Superadditivity is merely due to the V-condition being below baseline. We also found superadditivity within a small cluster of voxels in the left occipital pole (Fig. 4, panel 3) and the right supramarginal gyrus in the parietal cortex (not shown). Most remarkable, however, was that both MGN clusters survived the statistical threshold, displaying significant effects

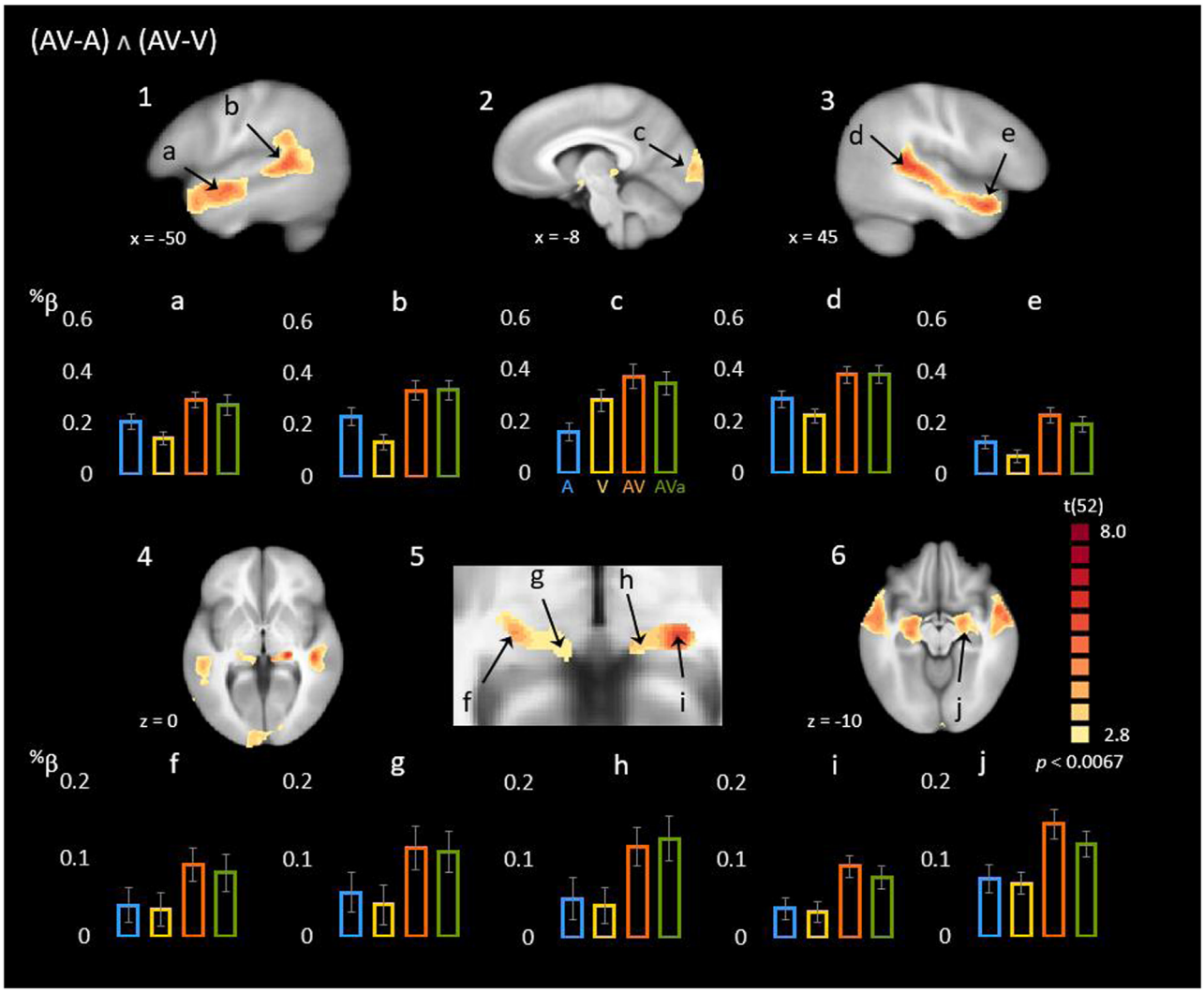


**Fig. 3.** Statistical Conjunction of the (AV-A) and the (AV-V) contrasts (Max. criterion).
Maps 1-6 show voxels with significant t-scores of the conjunction of the (AV-A) and (AV-V) contrasts FDR- corrected (q = 0.05) for multiple comparisons. Bar graphs represent selected % transformed predictor values for A, V, AV and AVa conditions averaged over 4 functional voxels centered around peak voxel locations (see Table 3). a) Left anterior superior temporal gyrus; b) left pSTS; c) left occipital pole; d) right pSTS; e) left anterior STS; f) left LGN; g) left MGN; h) right MGN; i) right LGN; j) right amygdala.

**Table 4**
Clusters of significant activity resulting from the subtraction of the sum of the predictor values for the A and V conditions from the AV condition (superadditivity). Significant clusters are numbered and reported with their t-statistic and location in Talairach space in the order of cluster size. In cases where clusters spanned over more than one anatomical or functional region additional peak voxels are reported together with their corresponding anatomical region.

| **AV > (A+V)** Cluster | | L/R | *t*-statistic | x | y | z | Voxels |
|---|---|---|---|---|---|---|---|
| 1 | Heschl's gyrus | R | 5.92 | 58 | -11 | 7 | 1244 |
| | Superior temporal sulcus | R | 5.89 | 52 | -9 | -10 | |
| 2 | Heschl's gyrys | L | 4.94 | -54 | -15 | 9 | 574 |
| | Superior temporal sulcus | L | 4.2 | -58 | -3 | -5 | |
| 3 | Insula (Wernicke) | L | 3.86 | -50 | -34 | 19 | 27 |
| 4 | Lingual gyrus (occipital pole) | L | 3.90 | -5 | -98 | -3 | 26 |
| 5 | STG (temporal pole) | R | 3.67 | 42 | 18 | -19 | 24 |
| 6 | Inf. Occipital gyrus | L | 3.43 | -21 | -89 | -9 | 24 |
| 7 | Medial geniculate nucleus | R | 3.54 | 17 | -23 | -3 | 22 |
| 8 | Medial geniculate nucleus | L | 4.1 | -15 | -23 | -4 | 22 |

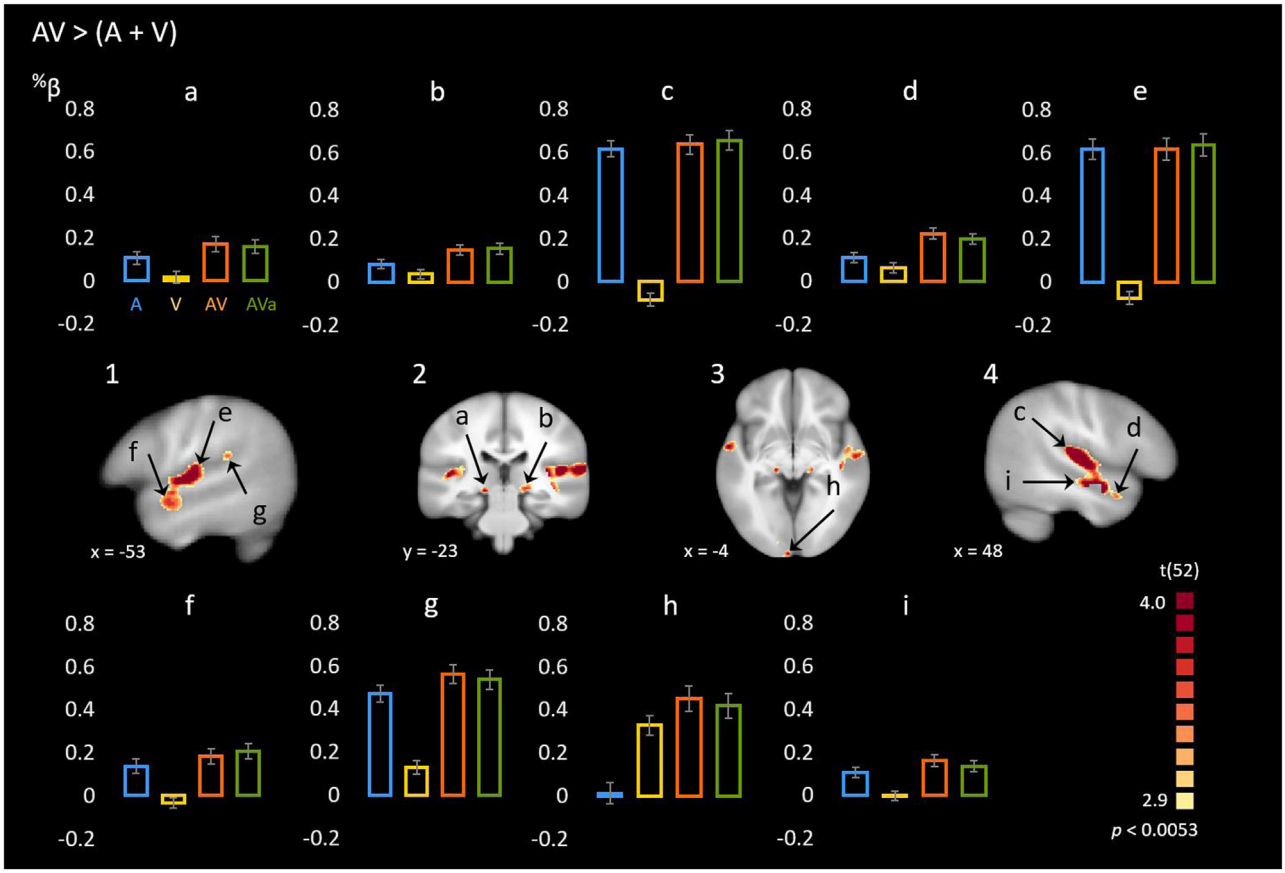


**Fig. 4.** Statistical comparison between AV and (A + V) (Superadditivity).
Maps 1-4 show voxels with significant t-scores of the comparison between the sum of the predictor values of the A and V conditions (A+V) and the AV condition FDR- corrected (q = 0.05) for multiple comparisons fulfilling the superadditive criterion. Bar graphs represent selected % transformed predictor values for A, V, AV and AVa conditions averaged over 4 functional voxels centered around peak voxel locations (see Table 4). a) Left MGN; b) right MGN; c) right pSTS; d) right anterior STS; e) left Heschl's gyrus; f) left anterior STS g) left pSTS; h) right occipital pole; i) right STS.

**Table 5**
Clusters of significant differences between the synchronous (AV) and asynchronous (Ava) audiovisual conditions.

| **AV vs. AVa** Cluster | | L/R | *t*-statistic | x | y | z | Voxels |
|---|---|---|---|---|---|---|---|
| 1 | Parietal lobe | R | -6.24 | 52 | -48 | 45 | 203 |
| 2 | Middle frontal gyrus | L | -5.26 | -36 | -8 | 47 | 51 |
| 3 | Superior frontal gyrus | R | -5.22 | 37 | 20 | 48 | 50 |

(Fig. 4, panels 2 and 3). We did not find evidence for superadditivity in the posterior STS.

### 3.5. Audiovisual (AV) versus asynchronous audiovisual (AVa)

We found three smaller clusters (Fig. 5, panels 1 and 2) in which BOLD was higher in the asynchronous condition, one in the right parietal cortex, one in the right prefrontal cortex and one in the left premotor cortex (Table 5). Closer inspection of the betas revealed that for both loci in the right hemisphere both conditions were below baseline (Fig. 5, bar graphs a and c). We did not find evidence for relatively increased BOLD activity for the synchronous condition. We therefore concluded that this experimental manipulation did not result in significant differences that are interpretable in regard to our hypotheses.

### 3.6. A - V

As described in the methods section, we used this contrast as an inclusion criterion for our sample because it was a more sensitive indicator of BOLD responses to the A and V stimuli than comparing unisensory responses to baseline. We also considered that under the likely assumption that in our experimental design the lack of rest conditions between blocks of stimulation could result in a "high" baseline, some unisensory effects would not be observable when comparing to baseline alone.

As expected, the respective A and V conditions resulted in much of the same regional activations they did when compared to baseline (Appendix, Supplementary Figure 1, Table 1). However, several differences are worth noting here. First, and to our initial surprise given the results from the A alone analysis, the A condition resulted in significantly higher BOLD response in much of the visual association cortex with centers of gravity in the bilateral lingual gyrus (Appendix, Supplementary Figure 1, panel 3). This cluster extended dorsally into the parietal cortex (Appendix, Supplementary Figure 1, panels 7 and 8) around the midline and included the postcentral gyrus. A look at the percent signal change from an ROI (Appendix, Supplementary Figure 1, panel 3) at the centers of gravity revealed that all conditions containing a visual stimulus but not the A- condition were significantly below baseline at these locations (Appendix, Supplementary Figure 1, panel 3 and associated bar graph).

Another remarkable observation was that this contrast revealed activations along the subcortical auditory pathway (Appendix, Supplementary Figure 1, panel 2). The BOLD response in the bilateral medial geniculate nuclei to the A condition was significantly larger than in the

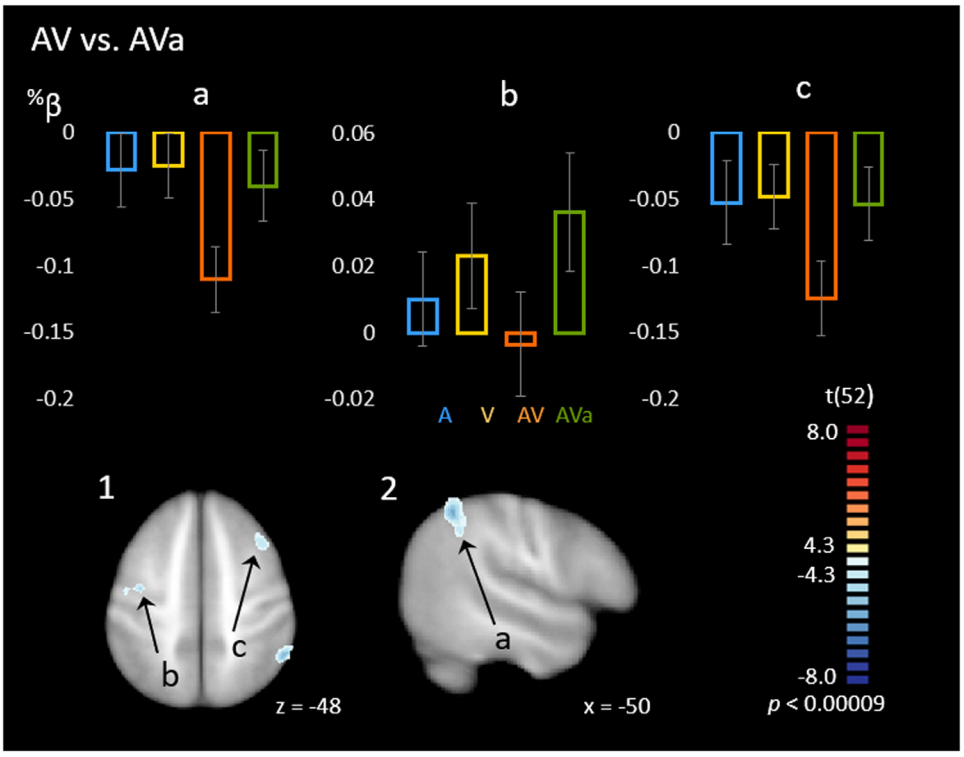


**Fig. 5.** Statistical comparison between AV and Ava conditions.
Maps 1 and 2 show voxels with significant t-scores of the comparison of the AV (red) and AVa (blue) conditions FDR- corrected (q = 0.05) for multiple comparisons. Bar graphs represent selected % transformed predictor values for A, V, AV and AVa conditions averaged over 4 functional voxels centered around peak voxel locations (see Table 5). a) Right parietal lobe; b) left SFG; c) left middle frontal gyrus.

V-condition (and larger than any of the conditions containing a visual stimulus). The same pattern was observed in both inferior colliculi (Appendix, Supplementary Figure 1, panel 6) although we interpret this with caution due to the small size of these structures and the inability to achieve perfect anatomical matching between subjects. We also found two clusters (A > V) in the bilateral crus cerebri (Appendix, Supplementary Figure 1, panel 2).

### 3.7. Conjunction of auditory and visual conditions (A ∧ V)

We conducted a conjunction analysis to determine which regions of the brain exhibited BOLD responses that were significantly above baseline for both auditory and visual conditions (Appendix, Supplementary Figure 2, Table 2). We found this to be the case in the right pSTS/G (Appendix, Supplementary Figure 2, panel 3) and the ATL (Appendix, Supplementary Figure 2, panel 1), IFG and precentral gyrus in the left hemisphere (Appendix, Supplementary Figure 2, panel 2). We also found the left lentiform nucleus to exhibit a significant response to unisensory A and V stimuli (not shown).

### 3.8. Behavioral task results

In line with previous findings (Ross et al., 2011) (Sumby, 1954) the RM-ANOVA returned main effects (Greenhouse-Geisser corrected for violation of sphericity) of SNR [$F_{(4.78, 239)} = 16.7$; $p < 0.001$; $\eta_P^2 = .25$] and condition [$F_{(1, 50)} = 8.93$; $p = 0.004$; $\eta_P^2 = .152$] showing that performance decreased as SNR decreased and was significantly better when visualized speech was present (see Fig. 6 and Table 6). Audiovisual gain showed the characteristic inverted- u shape relationship to SNR that we have reported in the past (Foxe et al., 2015; Ma et al., 2009; Ross et al., 2015; Ross et al., 2011; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007; Ross, Saint-Amour, Leavitt, Molholm, et al., 2007) with a maximum ($M = 37.12\%$; $SD = 18.7\%$) at intermediate (-9dB) intelligibility. We found no significant interaction between both factors indicating that SNR affected performance in the A and AV condition in a similar manner [$F_{(4.74, 237)} = 1.72$; $p = 0.136$; $\eta_P^2 = .03$]. Neither age [$F_{(1, 50)} = 3.66$; $p = 0.061$; $\eta_P^2 = .068$] nor biological sex [$F_{(1, 50)} = 0.014$; $p = 0.907$; $\eta_P^2 < 0.001$] were significant.

Overall, participants were able to speechread the correct word in $M = 13.65\%$ ($SD = 9.61$) of cases with no appreciable difference between males ($M = 12.57\%$; $SD = 8.09\%$) and females ($M = 14.86\%$; $SD = 11.13\%$) ($F_{(1, 50)} = 0.516$; $p = 0.476$; $\eta_P^2 = 0.01$) and no effect of age ($F_{(1, 50)} = 0.274$; $p = 0.603$; $\eta_P^2 = 0.005$) (Table 7). We found no relationship between performance in the auditory condition at low SNRs with speechreading performance $r(51) = 0.013$, p = 0.926 and a positive relationship with AV performance at low SNRs $r(51) = 0.33$, $p = 0.015$.

### 3.9. Association between BOLD responses and behavioral performance

#### 3.9.1. Audiovisual gain

This analysis was performed to identify brain regions that are involved in the gain conferred by the presence of congruous visual input (Fig. 6). For this we conducted voxel-wise correlations between beta weights of the AV-A contrast and the difference between audiovisual (AV) and auditory alone (A) performances in the behavioral experiment. We used the map of the voxel-wise *Pearson r* statistic of the AV-A contrast, thresholded at $p = 0.01$ as the input for the Monte Carlo cluster estimation which, after 5000 iterations returned a cluster threshold of 74 voxels. The resulting map showed left hemispheric clusters of significant positive correlations in the primary visual cortex ($r(51) = 0.469$; $p < 0.001$), the cuneus ($r(51) = 0.46$; $p = 0.001$) and the posterior middle temporal gyrus ($r(51) = 0.455$; $p = 0.001$).

#### 3.9.2. Audiovisual

Cluster threshold estimation was carried out on the r- map reflecting significant correlations between the AV BOLD response and AV performance in the speech in noise task thresholded at $p = 0.01$. Only a cluster in the inferior parietal lobe survived a threshold of 67 voxels ($r(51) = 0.36$; $p = 0.008$).

#### 3.9.3. Visual alone (Speechreading)

The correlation map between BOLD response to the V-condition and performance in the speechreading condition was thresholded at $p = 0.01$

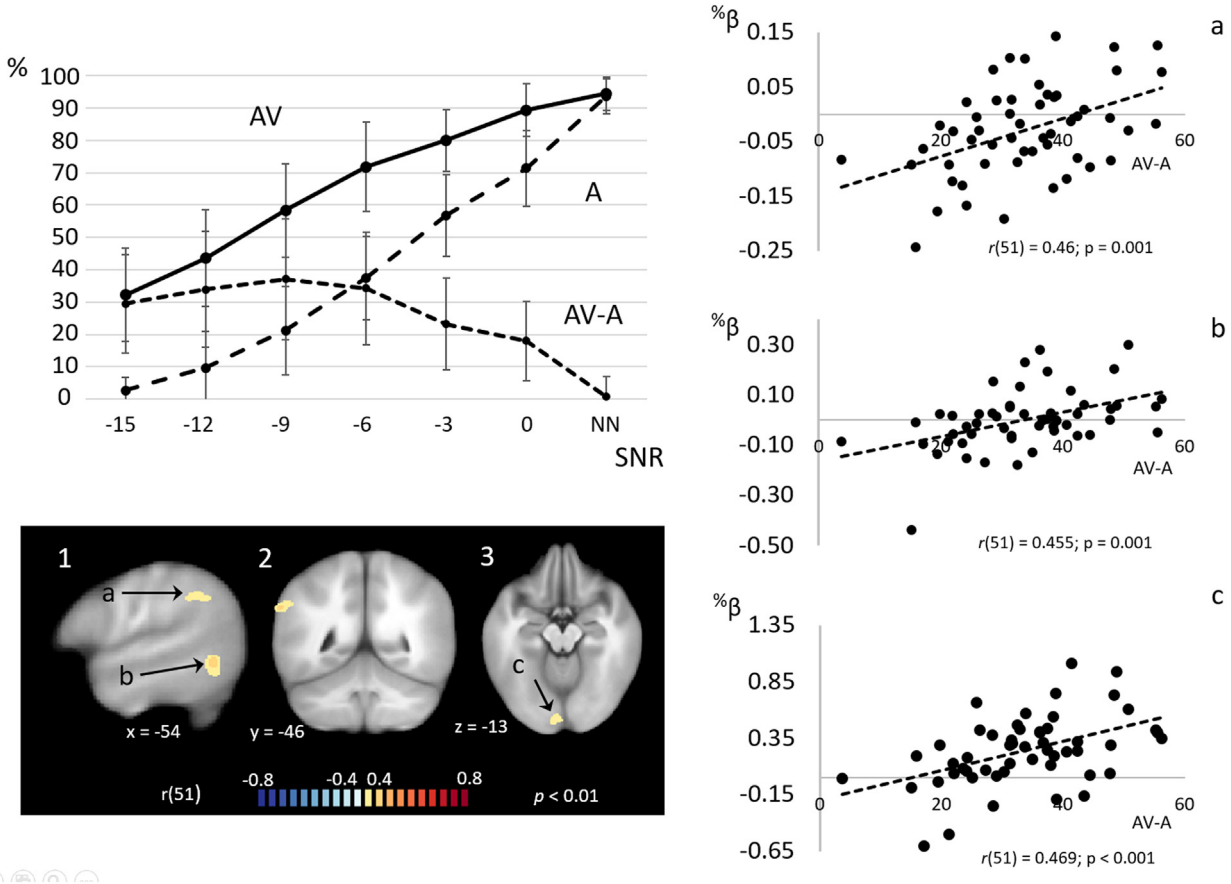


**Fig. 6.** Performance in the behavioral task and associations with brain activity.
Line graph represents % correct performance in the A and AV conditions as well as audiovisual gain (AV-A) over seven SNR conditions with error bars representing standard deviations from the mean. Maps 1-3 represent significant Pearson r correlation coefficients with an applied cluster correction of 74 functional voxels with clusters in the a) cuneus; b) pMTG and c) occipital cortex of the left hemisphere. Scatter plots show % transformed predictor values (AV-A) (y-axis) for each participant averaged over 4 voxels at the centers of the clusters shown in the statistical map in relationship to behavioral audiovisual gain (AV-A) on the x- axis.

**Table 6**
ANOVA table. F-test of the effects of Condition, SNR, Sex and Age on speech perception.

| F-test of the effects of Condition, SNR, Sex and Age | | | | | | |
|---|---|---|---|---|---|---|
| Source | SS | df | MS | F | *p* | $\eta_P^2$ |
| *Tests of Within-Subjects Effects and interactions* | | | | | | |
| Condition | 1670.83 | 1 | 1670.83 | 8.93 | 0.004 | 0.15 |
| Error Condition | 9354.81 | 50 | 187.1 | | | |
| SNR | 8765.89 | 6 | 1830.55 | 16.69 | 0.001 | .25 |
| Error SNR | 26249.95 | 300 | 87.5 | | | |
| Condition x Age | 58.46 | 1 | 58.46 | 0.31 | 0.58 | 0.006 |
| Condition x Sex | 387.89 | 1 | 387.89 | 2.07 | 0.16 | 0.04 |
| SNR x Age | 731.03 | 4.79 | 152.66 | 1.39 | 0.23 | 0.027 |
| SNR x Sex | 1704 | 4.79 | 355.96 | 3.25 | 0.008 | 0.067 |
| *Tests of Between-Subjects Effects* | | | | | | |
| Age | 62178.4 | 1 | 62178.4 | 132.3 | <0.001 | 0.726 |
| Sex | 6.49 | 1 | 6.49 | 0.14 | 0.91 | 0 |
| Error | | 1 | | | | |

and submitted to the cluster threshold estimation resulting in a minimum cluster size of 69 voxels. No cluster in the map survived this threshold.

### 3.9.4. Auditory alone

We computed whole brain voxel-wise correlations between the BOLD predictor of the A- condition and the average performance (% correct) in the auditory condition of the speech in noise behavioral task. The correlation map was thresholded at $p = 0.01$ and submitted to the Monte Carlo cluster estimation procedure which returned a cluster threshold of 62 voxels. No plausible significant correlations between the BOLD response and behavioral performance were found (a cluster in the cerebellum was driven by two outliers).

**Table 7**
ANOVA table. F-test of the effects of Sex and Age on speechreading performance.

| F-test of the effects of Sex and Age on speechreading performance | | | | | | |
|---|---|---|---|---|---|---|
| Source | SS | df | MS | F | $p$ | $\eta_P^2$ |
| *Tests of Between-Subjects Effects* | | | | | | |
| Age | 25.78 | 1 | 25.78 | 0.27 | 0.6 | 0.005 |
| Sex | 48.65 | 1 | 48.65 | 0.52 | 0.48 | 0.01 |
| Error | 4710.53 | 50 | 94.21 | | | |

### 3.9.5. Questionnaire performance

Forty eight out of 53 participants completed the 10-item questionnaire which can be found in the appendix. Three of the participants had prior knowledge of the story. A majority of seventy five percent of the participants answered six or more questions correctly and the average number of correct answers was 7.6 ($SD = 2.2$). Due to the interspersed V-alone blocks, it was fully expected that task performance would remain below a perfect score.

## 4. Discussion

The goal of this fMRI study was to investigate brain regions showing audiovisual enhancement during perception of narrative speech. We expected this enhancement to be evident in regions previously identified as part of the MS speech network (Erickson et al., 2014) (Calvert & Thesen, 2004) but also in regions downstream from known MSI sites reflecting the effects of successful integration in the larger speech processing network rather than MSI alone. We reasoned that if AV integration results in improved perception of the auditory speech signal, then the consequences of integration should be observable in the BOLD response in regions underlying the perception and semantic processing of the respective speech stimulus at word, sentence and narrative levels. Further, it may be possible, depending on the content, to observe enhancing effects on other cognitive functions such as memory retrieval and emotional processing. Since most studies investigating MS integration are interested in the processes and regions underlying MS modulation and convergence, these effects have been considered confounds with the intention to eliminate them through experimental control. However, this deprives us of the observation of broader effects that AV integration might have on the speech and language processing network.

The current study revealed an extensive network of MS enhancement. This network included well established sites of MS integration as well as parts of the semantic language network. We also found enhancement in the primary visual cortex and the bilateral amygdalae, and extralinguistic regions not usually associated with MS integration. Finally, our analysis revealed involvement of thalamic brain regions along the visual and auditory pathways more commonly associated with early sensory processing.

### 4.1. Unisensory conditions

Before we assessed MS enhancement, we explored how the unisensory narrative speech stimulus engaged the speech network. At the word level, core perisylvian language areas are active (Binder, 1997; Binder et al., 2000; Hertrich et al., 2020). Sentences and narratives differ in several regards from more simple speech stimuli, such as words or phonemes that have been used in most previous studies of audiovisual speech processing, and are known to involve core perisylvian language regions. The addition of syntactic and semantic information at the sentence level involves additional perisylvian cortex "spreading" along the posterior STS/G into the ATLs (Ardila et al., 2016; Binder, 2017; Price, 2012). Narratives contain additional, more complex semantic information in the form of thematic content tying information delivered over multiple sentences into a common overarching context (Hertrich et al., 2020; Xu et al., 2005; Xu et al., 2017). Further, processing at discourse level requires the listener to extract meaning increasingly tying lexical information to world knowledge creating mental representations of the narrative (Xu et al., 2005) and their potential social implications. It is reasonable to assume that this places additional demands on attention, working memory and higher cognitive functions such as theory of mind and may evoke emotions and visual imagery. Therefore, the increasing complexity of the language stimulus involves a multitude of extralinguistic cognitive operations involving extrasylvian regions that are reflected in the BOLD signal (de Heer et al., 2017; Huth et al., 2016; Lerner et al., 2011). This increase in complexity has been shown to engage left frontal regions (Lerner et al., 2011) supported by evidence from studies in frontotemporal dementia (Ash et al., 2006; Peelle & Grossman, 2008) showing that damage to these regions particularly affects discourse level processing.

In the A-condition we found bilateral activation of perisylvian regions along the superior temporal plane (Fig. 1, panels 1 and 4) and bilateral engagement of the articulatory motor and supplementary motor cortex but activity in the IFG and dlPFC was left lateralized. It has been claimed that naturalistic language stimuli lead to more bilateral cortical engagement (Hamilton & Huth, 2020; Jung-Beeman, 2005).

Our findings add to the mounting evidence that the motor cortex is engaged in speech perception (Heyes & Catmur, 2022; Schomers & Pulvermuller, 2016; Scott et al., 2009; Wilson et al., 2004) (Pulvermuller & Fadiga, 2010) (Cogan et al., 2014) which appears to show enhancement during speech perception in noise (Nuttall et al., 2017, 2018). It has been suggested that these findings reflect the action of a mirror neuron system (Iacoboni, 2008; Meister et al., 2007; Pulvermuller et al., 2006) and thereby a possible mechanism for a language perception module as postulated by Liberman (Liberman & Mattingly, 1985; Rizzolatti & Arbib, 1998). However, a crucial prediction of this theory is that speech perception and speech motor action share the same neural substrate. More recent lesion studies have cast doubt on this notion showing that patients with impaired speech production can still show unimpaired speech perception (Hickok et al., 2011; Rogalsky & Hickok, 2011; Stasenko et al., 2015). Using intracranial recordings over perisylvian cortex while human participants listened and spoke Cheung et al. (Cheung et al., 2016) found that activity over the motor cortex was substantially different during speech perception than speech production of the same sounds. Interestingly, the pattern of activity during listening was organized along acoustic features similar to the auditory cortex while speaking was organized along articulatory features. This suggests that speech perception and production recruit different networks within the motor cortex. The precise role of the motor cortex for speech perception and under which conditions it makes a necessary contribution to it remain under investigation.

An interesting finding was that the V condition (V > 0) appeared to engage an extended network of left frontal regions including the IFG (Fig. 2, Panel 1). One possible explanation is that participants subvocalize during speechreading or are engaged in cognitive processes that evoke semantic activity. On the other hand, the contrast between the unisensory conditions (A vs. V) did not result in higher activations to the A condition in these same regions as one would expect (Appendix, Supplementary Figure 1, panel 1), since these frontal regions are involved in operations resulting from narrative speech processing (Binder et al., 2009; Hertrich et al., 2020; Xu et al., 2005).

We made several further novel observations when contrasting the A and V conditions directly (A vs. V). First, it was apparent that the bilateral MGNs are engaged during listening to natural narrative speech (Appendix, Supplementary Figure 1, panel 2). This is expected, given that their function as thalamic relays along the auditory pathway is well known. However, effects in these subcortical structures are rare in fMRI experiments due to their small size and accompanying issues that will be discussed further below.

We also observed that activation in much of the visual association cortex was larger to the auditory alone stimulus than to the visual articulation, especially in the bilateral lingual gyrus (Appendix, Supplemen-

tary Figure 1, panel 3 and 7). All three conditions containing a visual stimulus are far below baseline whereas the auditory stimulus was at baseline (Appendix, Supplementary Figure 1, bar graph). This pattern is reversed in the nearby primary visual cortex at the occipital pole (Fig. 2, bar graph c). We speculate that activity in visual association cortex to the auditory stimulus is due to visual imagery evoked by the story narrative (Bergen et al., 2007; Hertrich et al., 2020) (Pearson, 2019) that is presumably absent in the V condition. We further speculate that this activity is suppressed in AV conditions when a visual stimulus is present because of a shift of attention to the visual stimulus (i.e. the speaker). This is an incidental finding not related to the original purpose of the study, the investigation of audiovisual enhancement in narrative processing, but is nevertheless interesting and relevant to report because it may shed light on a mechanism related to evoked visual imagery during auditory stimulation and its suppression.

### 4.2. Audiovisual enhancement

We used the following conjunction approach [(AV-A) ∧ (AV-V)] to identify regions of MS enhancement. This was largely satisfied for regions along the left and right STS (Fig. 3, panels 1 and 3) and included posterior sections of the STS commonly associated with MSI (Beauchamp et al., 2004).

The bilateral sections of the STS anterior of regions typically associated with MSI are well known to be part of the semantic system (Binder et al., 2009; de Heer et al., 2017; Hickok et al., 2018) and these findings provoke the question why these regions are enhanced by a MS stimulus. One explanation could be that despite our efforts to make the stimulus intelligible under unisensory auditory stimulation, the additional information from visual articulation resulted in an increase in intelligibility which in turn affected the content processed by the semantic system. If this was the case, however, this increase in intelligibility would likely be evident in modality specific auditory regions in the superior temporal plane. We did not find evidence for a significantly higher BOLD effect in the AV condition than in the A condition (Fig. 1, bar graphs b and d) due to possible ceiling effects in these regions (also see discussion on superadditivity). Responses in the STS to the unisensory conditions were overall lower, leaving “room” for MS enhancement.

Another possible explanation is that the MS stimulus is inherently more salient than unisensory stimuli alone. Moreover, more than just articulatory features used for linguistic analysis, the MS stimulus conveys important non-linguistic contextual information through tone, timing and volume of the voice, facial expressions, posture and head movement (Munhall & Buchan, 2004; Munhall & Johnson, 2012). In a natural conversation this additional information may be used by the speaker to aid the delivery of information, clarify intent and project emotional state. In the case of the reading of a story by a trained actress, as was the case in our experiment, this MS contextual information is more complex because the speaker does not deliver her own state or intent but that of the characters and their roles in the story. Given the complexity of a natural narrative, it is apparent how this non-linguistic contextual information renders the MS stimulus particularly salient. The notion of a widespread, non-specific effect of saliency of the MS stimulus is supported by our finding of MS enhancement in the bilateral amygdalae. Neither visual speech articulation and emotional facial expression nor listening to the auditory narrative with its emotional content was sufficient to engage the amygdalae compared to baseline.

MS enhancement was also observed in the primary visual cortex around the occipital poles (Fig. 3, panel 4). Less prevalent but significant activity was also found in higher-order areas considered part of the ventral visual pathway in inferior temporal regions of the left hemisphere. These regions correspond well with regions previously identified as involved in visual speech perception (Bernstein & Liebenthal, 2014). Thus, MS speech integration is not simply associated with visual influences on auditory processing, but rather, there is a clear bi-directionality to these influences, with substantial modulation of visual processing seen as a result of auditory inputs. This corresponds well with findings from human intracranial studies by our and other groups where auditory inputs significantly impacted visual cortical processing of simultaneously presented visual inputs across a substantial extent of visual cortex (Brang et al., 2022; Brang et al., 2015; Mercier et al., 2013; Mercier et al., 2015).

### 4.3. Subcortical audiovisual enhancement

Perhaps the most striking finding of this study is that of focal enhancements in the posterior thalamus involving the medial and lateral geniculate nuclei and the pulvinar (Fig. 3, panel 2, 4 and 5). That subcortical structures such as the superior colliculus (Wallace et al., 1998; Xu et al., 2014; Yu et al., 2013), inferior colliculus (Gruters & Groh, 2012) and some of the thalamic nuclei, especially the medial pulvinar (Cappe et al., 2009; Dietrich et al., 2013; Froesel et al., 2021), are involved in MS processing is now well-known. However, in the past these structures have been investigated in regard to low- order sensory processes using relatively simple stimuli and are therefore rarely associated with audio-visual speech processing (although see (Hebb & Ojemann, 2013)).

A spate of neuroimaging studies has indeed pointed to such subcortical MS processing under a variety of conditions. For example, Noesselt and colleagues (Noesselt et al., 2010) showed that functional connectivity between both the medial and lateral geniculate nuclei with their respective sensory cortices, as well as with the STS, was modulated under MS conditions and that the strength of these couplings across participants was associated with performance on a visual stimulus detection task for difficult-to-detect low-contrast visual inputs. Using a MS target detection task where synchronized auditory “pips” have been found to substantially improve target detection in cluttered moving visual scenes (Van der Burg et al., 2008), van der Burg and colleagues asked if variance in this MS ability could be associated with both structural and functional connectivity between thalamic nuclei and sensory-cortical representations. Using diffusion tensor imaging and probabilistic tractographic techniques, they asked whether connectivity between task-specific auditory and visual cortex (A1 and V4) and in turn, between these regions and their respective thalamic nuclei, would predict inter-individual differences in MS target detection. They found that the strength of structural connectivity between the cochlear nucleus, the medial geniculate body and primary auditory cortex was related to this integrative ability.

Perhaps more directly relevant to the current work is evidence from studies using speech stimuli showing MS responses in the brainstem. Fairhall and Macaluso (Fairhall & Macaluso, 2009) showed that attention to congruent AV speech stimuli resulted in increased activation in the superior colliculus compared to attention to incongruent stimuli. In an electrophysiological study, Musacchia et al. (Musacchia et al., 2006) recorded the auditory brainstem response (ABR) while participants listened to synthesized phonemic stimuli (e.g. /da/). There were three different conditions, one with no visual input where only the phonemes were heard, one where phonemes were accompanied by either congruent or incongruent visual articulations, and one where only the visual tokens were presented during silence. They found modulation of both latency and amplitude of the auditory brainstem response (ABR) under audio-visual conditions, effects that began as early as 11 ms following acoustic input, and these MS effects were also found to differ as a function of congruence between the visual and acoustic phonemic inputs. The work suggests, as do our results here, that ongoing visual articulatory inputs can shape the auditory system’s response to anticipated acoustic inputs, and that this top-down modulatory effect can be instantiated extremely early in the subcortical processing hierarchy – indeed, even before auditory information reaches the relevant thalamic nuclei. The authors hypothesize that this effect may reflect a cortical gating or attentional modulation mechanism and name the corticofugal system as a possible physiological candidate. There is now mounting evidence

that this complex system is not limited to the auditory pathway, with effects of attention to reaching down to the cochlea and auditory pathway (see review by (Elgueda & Delano, 2020)) allowing for more complex interactions from receptor to cortical levels.

It is of interest to note that recordings directly in rat auditory thalamus have shown that visual inputs can substantially modulate the early phase of auditory thalamic responsivity, significantly impacting behavior in these animals (Komura et al., 2005).

### 4.4. Superadditivity

The comparison (AV > A + V) in fMRI is largely adopted from animal electrophysiology studies that have shown neurons exhibiting stronger responses to MS as opposed to unisensory stimulation (Stein & Stanford, 2008; Xu et al., 2014). The rationale for adopting this method for BOLD fMRI that reflects activity from large populations of neurons was that BOLD activation is a time invariant-linear system where activation to two stimuli presented together is equivalent to the sum of the two stimuli presented individually (see James (James, 2012) for a review). If a region contains MS cells, the evoked activity to a MS stimulus is predicted to exceed the sum of the unisensory responses.

In practice this theoretical model to identify MS regions has proven to be too conservative, with many fMRI studies failing to show superadditivity in regions well known to be involved in MSI (Beauchamp, 2005; James, 2012). In the present study it was our intention to explore a network of regions showing MS enhancement beyond the sites typically reported to be involved in MS integration *per se*, so we adopted the less conservative max criterion (Altieri et al., 2011; Beauchamp, 2005). We conducted an additional analysis of MS enhancement using the additive criterion with the goal to explore whether this criterion would isolate a reduced set of classic MS integration sites, expecting these regions to overlap with clusters identified with the max criterion.

The results of this analysis highlighted some of the problems associated with this criterion. First, we failed to replicate superadditivity in the posterior STS (Fig. 4, panels 1 and 4). In the temporal lobes, superadditivity is apparent in Heschl's gyrus and the superior temporal plane in both hemispheres covering large parts of the auditory cortex but only in the more anterior STS (Fig. 4, panels 1 and 4). In the auditory cortex, the effect is mainly due to the fact that V is below baseline (Fig. 4, bar graphs c and e). The difference between AV and A is not significant. If one assumes that brain activity in the auditory cortex as measured as BOLD effects reflects the quality of the perceptual effect then it would be hard to argue that the AV condition would convey any perceptual advantage over the A condition. We attribute the lack of difference between the A and AV conditions to the high intelligibility of the auditory stimulus with the result of a ceiling effect.

Therefore, in conditions of high intelligibility and considering the somewhat arbitrary nature of the baseline, superadditive effects can be misleading. Nevertheless, we found genuine superadditive effects in the bilateral MGNs (Fig. 4, panels 2 and 3), the anterior portions of the STS (Fig. 4, panels 1 and 4) and a small cluster in the left occipital pole (Fig. 4, panel 3).

### 4.5. MS temporal congruency

One way to overcome the inherent difficulty of identifying MS regions by comparing MS to unisensory BOLD responses (James, 2012; Stevenson et al., 2009) is to adopt an experimental approach that allows for comparison of two MS conditions to one another that engage MS regions differentially. The approach adopted here that was successfully used in the past (Miller & D'Esposito, 2005; Stevenson et al., 2010; van Atteveldt et al., 2007; van Wassenhove et al., 2007) was to offset the auditory and the visual tracks sufficiently to prevent an integration of sound and visual articulatory movements. Based on the extant literature, a 400 ms delay of the visual signal appeared appropriate for this purpose. Participants are able to detect an asynchrony of an audiovisual speech signal at a 132ms delay of the visual signal (Dixon and Spitz, 1980) see also (van Wassenhove et al., 2007). The strength of the McGurk effect is reliably different at a 60ms delay of the visual signal (Munhall & Buchan, 2004; Munhall et al., 1996). In an fMRI study by Stephenson et al. (Stevenson et al., 2010), a 400 ms offset (visual lead) was effective in generating BOLD differences between synchronous and asynchronous audiovisual stimulus material.

Based on previous findings (Marchant et al., 2012; Noesselt et al., 2007; Stevenson et al., 2010) (Okada et al., 2013)we expected the temporal congruency of the auditory and visual speech inputs to impact the degree to which the MS network was engaged, particularly in the posterior superior temporal cortex.

Much to our surprise, our experimental manipulation was not effective in evoking the expected effects and attempts at an explanation must remain speculative. First, an offset of 400ms was not sufficient to prevent integration from taking place despite all previous evidence for reasons that might be specific to our stimulus material. It is also possible that over the course of the experiment, participants adapt to the asynchrony and thereby integrate the auditory and visual stimulus over a 400ms offset as part of a learning effect (Crosse et al., 2015; Luo et al., 2010). Another possible explanation is that in the asynchronous condition groups of MS neurons are engaged despite the lack of synchronicity and thus drive a BOLD response that is comparable to the synchronous AV condition. The activity of this group of neurons may not reflect a response to congruous auditory and visual information and would not result in MS enhancement under more degraded listening conditions. Since our stimuli were sufficiently intelligible, this activity may have had no detrimental effect on the perception of the auditory stimulus and therefore did not result in a difference in the BOLD signal between MS conditions.

### 4.6. Correlation with behavioral multisensory speech-in-noise task

The motivation for this analysis was to locate regions associated with performance in an audiovisual speech perception task. However, we would like to advise the reader to consider the results of this particular analysis as well as their interpretation as preliminary. According to a publication by Eklund et al., (Eklund et al., 2016) we were not able to meet sufficiently conservative criteria for the protection against false positives because the initial correlation maps before cluster threshold estimation did not exceed $p < 0.001$ for whole brain analysis (see methods section). We expected these regions to be part of the well-established perisylvian speech processing network. However, it was in fact the primary visual cortex, cuneus and the posterior middle temporal gyrus (pMTG) of the left hemisphere that showed significant association with MS gain in the behavioral task (Fig. 6, panels 1, 2 and 3). The involvement of the primary (V1) and secondary (cuneus) visual cortex suggests that the ability to benefit from visual articulation is associated with activity in the visual cortices and may reflect processes underlying the analysis of visual articulatory movement and/or attention to the visual stimulus (Vanni et al., 2001). There is strong evidence that the pMTG plays a key role in semantic cognition (Binder et al., 2009; Hoffman et al., 2012. In line with our finding lesion studies indicated this region in language comprehension at word level {Dronkers, 2004 #820; Liuzzi et al., 2020; Turken & Dronkers, 2011) and has been suggested to be part of the ventral stream of speech processing (Fridriksson et al., 2016; Hickok & Poeppel, 2004). Davey et al. (Davey et al., 2016) suggested that this structure integrates information related to more automatic aspects of semantic cognition (presumably associated with passive listening) often associated with the default mode network and effortful task-related semantic retrieval. This model fits well with the notion that during passive listening to a complex narrative in our experiment, semantic aspects of the DMN and task-related semantic retrieval are both engaged depending on the degree of effort required to follow the thematic content of the story. While we ensured

that the auditory stimulus in our experiment was sufficiently intelligible, it is likely that intelligibility varied over the course of the experiment and/or between subjects due to the difficulty to fully control this variable in a scanner environment causing a stimulus-dependent, flexible change of engagement of passive default mode and effortful task-dependent semantic retrieval. The ability to engage the pMTG might in turn be related to the ability to retrieve semantic information in the low-intelligibility context of our speech-in-noise experiment and therefore serve as a possible explanation for the correlation between pMTG BOLD signal and MS gain in the behavioral experiment. Finally, despite its intuitive appeal, we did not find support for the notion that individuals with difficulty perceiving degraded speech exhibit greater speechreading or greater audiovisual benefit under difficult listening conditions (Dias et al., 2021a). In fact, we found a positive association between A and AV conditions at low SNRs, the opposite of what is predicted under this hypothesis.

### 4.7. Conclusions

The current study, by using a naturalistic narrative stimulus set and imaging a substantially larger cohort than used in most previous studies, revealed a considerably more extensive network of MS enhancement. This network included "classic" sites of MS integration as well as parts of the semantic language network. We also found enhancement in extralinguistic regions not usually associated with MS integration, namely the primary visual cortex and the bilateral amygdalae. Analysis also revealed involvement of thalamic brain regions along the visual and auditory pathways more commonly associated with early sensory processing

We posit that under natural listening conditions, MS enhancement not only involves sites of MS integration but many regions of the wider semantic network and includes regions associated with extralinguistic perceptual and cognitive processing.

## Funding

Participant recruitment, phenotyping and neuroimaging/neurophysiology at the University of Rochester (UR) is conducted through cores of the UR Intellectual and Developmental Disabilities Research Center (UR-IDDRC), which is supported by a center grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (P50 HD103536 – to JJF). Neuroimaging and phenotpying at the Albert Einstein College of Medicine collaboration site was supported by the Rose F. Kennedy Intellectual and Developmental Disabilities Research Center (RFK-IDDRC), which is funded through a center grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (U54 HD090260 – to SM).

## Data sharing

The de-identified neuroimaging data and related materials are available publicly through the Dryad open-access repository of research data at the following URL: Ross, Lars (2022), LORAX, Dryad, Dataset, https://doi.org/10.5061/dryad.sj3tx967q under the digital object identifier (DOI): doi:10.5061/dryad.sj3tx967q.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Credit authorship contribution statement

**Lars A. Ross:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – review & editing. **Sophie Molholm:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition. **John S. Butler:** Resources, Data curation, Writing – review & editing. **Victor A. Del Bene:** Investigation, Data curation, Writing – review & editing. **John J. Foxe:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition.

## Data availability

The de-identified neuroimaging data and related materials are available publicly through the Dryad open-access repository of research data at https://doi.org/10.5061/dryad.sj3tx967q.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119598.

## References

Alexandrou, A.M., Saarinen, T., Kujala, J., Salmelin, R., 2020. Cortical entrainment: what we can learn from studying naturalistic speech perception. Lang. Cognit. Neurosci. 35 (6), 681–693. doi:10.1080/23273798.2018.1518534.

Alsius, A., Pare, M., Munhall, K.G., 2018. Forty years after hearing lips and seeing voices: the McGurk effect revisited. Multisens Res 31 (1-2), 111–144. doi:10.1163/22134808-00002565.

Altieri, N., Pisoni, D.B., Townsend, J.T., 2011. Some behavioral and neurobiological constraints on theories of audiovisual speech integration: a review and suggestions for new directions. Seeing Perceiv. 24 (6), 513–539. doi:10.1163/187847611X595864.

Ardila, A., Bernal, B., Rosselli, M., 2016. How localized are language brain areas? A review of brodmann areas involvement in oral language. Arch. Clin. Neuropsychol. 31 (1), 112–122. doi:10.1093/arclin/acv081.

Ash, S., Moore, P., Antani, S., McCawley, G., Work, M., Grossman, M., 2006. Trying to tell a tale: discourse impairments in progressive aphasia and frontotemporal dementia. Neurology 66 (9), 1405–1413. doi:10.1212/01.wnl.0000210435.72614.38.

Ayres, A.J., 1979. Sensory Integration and the Child. Western Psychological Services, Los Angeles.

Beauchamp, M.S., 2005. Statistical criteria in FMRI studies of multisensory integration. Neuroinformatics 3 (2), 93–113. doi:10.1385/NI:3:2:093.

Beauchamp, M.S., Lee, K.E., Argall, B.D., Martin, A., 2004. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41 (5), 809–823. http://www.ncbi.nlm.nih.gov/pubmed/15003179; http://ac.els-cdn.com/S0896627304000704/1-s2.0-S0896627304000704-main.pdf?_tid=67d08a7e-df90-11e3-ad77-00000aab0f26&acdnat=1400529886_ca4061b8081fc953e76843230df9def0.

Benoit, C., Mohamadi, T., Kandel, S., 1994. Effects of phonetic context on audio-visual intelligibility of French. J. Speech. Hear. Res. 37 (5), 1195–1203. http://www.ncbi.nlm.nih.gov/pubmed/7823561.

Bergen, B.K., Lindsay, S., Matlock, T., Narayanan, S., 2007. Spatial and linguistic aspects of visual imagery in sentence comprehension. Cogn. Sci. 31 (5), 733–764. doi:10.1080/03640210701530748.

Bernstein, L.E., Liebenthal, E., 2014. Neural pathways for visual speech perception. Front. Neurosci. 8, 386. doi:10.3389/fnins.2014.00386.

Binder, J.R., 1997. Neuroanatomy of language processing studied with functional MRI. Clin. Neurosci. 4 (2), 87–94. http://www.ncbi.nlm.nih.gov/pubmed/9059758.

Binder, J.R., 2017. Current controversies on Wernicke's area and its role in language. Curr. Neurol. Neurosci. Rep. 17 (8), 58. doi:10.1007/s11910-017-0764-8.

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb. Cortex 19 (12), 2767–2796. doi:10.1093/cercor/bhp055.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex 10 (5), 512–528. http://www.ncbi.nlm.nih.gov/pubmed/10847601; http://cercor.oxfordjournals.org/content/10/5/512.full.pdf.

Bolognini, N., Leo, F., Passamonti, C., Stein, B.E., Ladavas, E., 2007. Multisensory-mediated auditory localization. Perception 36 (10), 1477–1485. http://www.ncbi.nlm.nih.gov/pubmed/18265830.

Brandwein, A.B., Foxe, J.J., Butler, J.S., Frey, H.P., Bates, J.C., Shulman, L.H., Molholm, S., 2014. Neurophysiological indices of atypical auditory processing and multisensory integration are associated with symptom severity in Autism. J. Autism Dev. Disord. doi:10.1007/s10803-014-2212-9.

Brandwein, A.B., Foxe, J.J., Butler, J.S., Frey, H.P., Bates, J.C., Shulman, L.H., Molholm, S., 2015. Neurophysiological indices of atypical auditory processing and multisensory integration are associated with symptom severity in autism. J. Autism Dev. Disord. 45 (1), 230–244. doi:10.1007/s10803-014-2212-9.

Brandwein, A.B., Foxe, J.J., Russo, N.N., Altschuler, T.S., Gomes, H., Molholm, S., 2011. The development of audiovisual multisensory integration across childhood and early adolescence: a high-density electrical mapping study. Cereb. Cortex 21 (5), 1042–1055. doi:10.1093/cercor/bhq170.

Brang, D., Plass, J., Sherman, A., Stacey, W.C., Wasade, V.S., Grabowecky, M., Ahn, E., Towle, V.L., Tao, J.X., Wu, S., Issa, N.P., Suzuki, S., 2022. Visual cortex responds to sound onset and offset during passive listening. J. Neurophysiol. 127 (6), 1547–1563. doi:10.1152/jn.00164.2021.

Brang, D., Towle, V.L., Suzuki, S., Hillyard, S.A., Di Tusa, S., Dai, Z., Tao, J., Wu, S., Grabowecky, M., 2015. Peripheral sounds rapidly activate visual cortex: evidence from electrocorticography. J. Neurophysiol. 114 (5), 3023–3028. doi:10.1152/jn.00728.2015.

Callan, D.E., Jones, J.A., Munhall, K., Callan, A.M., Kroos, C., Vatikiotis-Bateson, E., 2003. Neural processes underlying perceptual enhancement by visual speech gestures. Neuroreport 14 (17), 2213–2218. doi:10.1097/01.wnr.0000095492.38740.8f.

Calvert, G.A., 2001. Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb. Cortex 11 (12), 1110–1123. http://www.ncbi.nlm.nih.gov/pubmed/11709482; http://cercor.oxfordjournals.org/content/11/12/1110.full.pdf.

Calvert, G.A., Brammer, M.J., Bullmore, E.T., Campbell, R., Iversen, S.D., David, A.S., 1999. Response amplification in sensory-specific cortices during crossmodal binding. Neuroreport 10 (12), 2619–2623. http://www.ncbi.nlm.nih.gov/pubmed/10574380.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., David, A.S., 1997. Activation of auditory cortex during silent lipreading. Science 276 (5312), 593–596. http://www.ncbi.nlm.nih.gov/pubmed/9110978.

Calvert, G.A., Campbell, R., Brammer, M.J., 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr. Biol. 10 (11), 649–657. http://www.ncbi.nlm.nih.gov/pubmed/10837246; http://ac.els-cdn.com/S0960982200005133/1-s2.0-S0960982200005133-main.pdf?_tid=52218f38-dae6-11e3-8299-00000aab0f6b&acdnat=1400017030_38c2719199bb33d2a39bad7c6778e415.

Calvert, G.A., Thesen, T., 2004. Multisensory integration: methodological approaches and emerging principles in the human brain. J. Physiol. Paris 98 (1-3), 191–205. doi:10.1016/j.jphysparis.2004.03.018.

Cappe, C., Morel, A., Barone, P., Rouiller, E.M., 2009. The thalamocortical projection systems in primate: an anatomical support for multisensory and sensorimotor interplay. Cereb. Cortex 19 (9), 2025–2037. doi:10.1093/cercor/bhn228.

Cheung, C., Hamiton, L.S., Johnson, K., Chang, E.F., 2016. The auditory representation of speech sounds in human motor cortex. Elife 5, e12577. doi:10.7554/eLife.12577, ARTN.

Cogan, G.B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., Pesaran, B., 2014. Sensory-motor transformations for speech occur bilaterally. Nature 507 (7490), 94. doi:10.1038/nature12935.

Coltheart, M., 1981. The MRC psycholinguistic database. Q. J. Exp. Psychol. 33A, 497–505.

Crosse, M.J., Butler, J.S., Lalor, E.C., 2015. Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J. Neurosci. 35 (42), 14195–14204. doi:10.1523/JNEUROSCI.1829-15.2015.

Crosse, M.J., Di Liberto, G.M., Lalor, E.C., 2016. Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term Crossmodal temporal integration. J. Neurosci. 36 (38), 9888–9895. doi:10.1523/JNEUROSCI.1396-16.2016.

Davey, J., Thompson, H.E., Hallam, G., Karapanagiotidis, T., Murphy, C., De Caso, I., Krieger-Redwood, K., Bernhardt, B.C., Smallwood, J., Jefferies, E., 2016. Exploring the role of the posterior middle temporal gyrus in semantic cognition: Integration of anterior temporal lobe with executive processes. Neuroimage 137, 165–177. doi:10.1016/j.neuroimage.2016.05.051.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical cortical organization of human speech processing. J. Neurosci. 37 (27), 6539–6557. doi:10.1523/JNEUROSCI.3267-16.2017.

Dias, J.W., McClaskey, C.M., Harris, K.C., 2021a. Audiovisual speech is more than the sum of its parts: Auditory-visual superadditivity compensates for age-related declines in audible and lipread speech intelligibility. Psychol. Aging 36 (4), 520–530. doi:10.1037/pag0000613.

Dias, J.W., McClaskey, C.M., Harris, K.C., 2021b. Early auditory cortical processing predicts auditory speech in noise identification and lipreading. Neuropsychologia 161, 108012. doi:10.1016/j.neuropsychologia.2021.108012, ARTN.

Diederich, A., Colonius, H., 2004. Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time. Percept. Psychophys. 66 (8), 1388–1404. http://www.ncbi.nlm.nih.gov/pubmed/15813202.

Dietrich, S., Hertrich, I., Ackermann, H., 2013. Ultra-fast speech comprehension in blind subjects engages primary visual cortex, fusiform gyrus, and pulvinar - a functional magnetic resonance imaging (fMRI) study. BMC Neurosci. 14, 74. doi:10.1186/1471-2202-14-74.

Ding, N., Simon, J.Z., 2014. Cortical entrainment to continuous speech: functional roles and interpretations. Front. Hum. Neurosci. 8, 311. doi:10.3389/fnhum.2014.00311, ARTN.

Dixon, N.F., Spitz, L., 1980. The detection of auditory visual desynchrony. Perception 9, 719–721. doi:10.1068/p090719.

Dronkers, N.F., Wilkins, D.P., Van Valin, R.D., Redfern, B.B., Jaeger, J.J., 2004. Lesion analysis of the brain areas involved in language comprehension. Cognition 92 (1-2), 145–177. doi:10.1016/j.cognition.2003.11.002.

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. In: Proc Natl Acad Sci U S A, 113, pp. 7900–7905. doi:10.1073/pnas.1602413113.

Elgueda, D., Delano, P.H., 2020. Corticofugal modulation of audition. Curr. Opin. Physiol. 18, 73–78. doi:10.1016/j.cophys.2020.08.016.

Erickson, L.C., Heeg, E., Rauschecker, J.P., Turkeltaub, P.E., 2014. An ALE meta-analysis on the audiovisual integration of speech signals. Hum. Brain Mapp. 35 (11), 5587–5605. doi:10.1002/hbm.22572.

Fairhall, S.L., Macaluso, E., 2009. Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. Eur. J. Neurosci. 29 (6), 1247–1257. doi:10.1111/j.1460-9568.2009.06688.x.

Foxe, J.J., Del Bene, V.A., Ross, L.A., Ridgway, E.M., Francisco, A.A., Molholm, S., 2020. Multisensory audiovisual processing in children with a Sensory Processing Disorder (II): speech integration under noisy environmental conditions. Front Integr Neurosci 14, 39. doi:10.3389/fnint.2020.00039, ARTN.

Foxe, J.J., Molholm, S., 2009. Ten years at the multisensory forum: musings on the evolution of a field. Brain Topogr. 21 (3-4), 149–154. doi:10.1007/s10548-009-0102-9.

Foxe, J.J., Molholm, S., Del Bene, V.A., Frey, H.P., Russo, N.N., Blanco, D., Saint-Amour, D., Ross, L.A., 2015. Severe multisensory speech integration deficits in high-functioning school-aged children with Autism Spectrum Disorder (ASD) and their resolution during early adolescence. Cereb. Cortex 25 (2), 298–312. doi:10.1093/cercor/bht213.

Foxe, J.J., Schroeder, C.E., 2005. The case for feedforward multisensory convergence during early cortical processing. Neuroreport 16 (5), 419–423. http://www.ncbi.nlm.nih.gov/pubmed/15770144.

Frens, M.A., Van Opstal, A.J., Van der Willigen, R.F., 1995. Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Percept. Psychophys. 57 (6), 802–816. http://www.ncbi.nlm.nih.gov/pubmed/7651805.

Fridriksson, J., Yourganov, G., Bonilha, L., Basilakos, A., Den Ouden, D.B., Rorden, C., 2016. Revealing the dual streams of speech processing, 113, pp. 15108–15113. doi:10.1073/pnas.1614038114.

Froesel, M., Cappe, C., Ben Hamed, S., 2021. A multisensory perspective onto primate pulvinar functions. Neurosci. Biobehav. Rev. 125, 231–243. doi:10.1016/j.neubiorev.2021.02.043.

Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15 (4), 870–878. doi:10.1006/nimg.2001.1037.

Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48 (1), 63–72. doi:10.1016/j.neuroimage.2009.06.060.

Gruters, K.G., Groh, J.M., 2012. Sounds and beyond: multisensory and other non-auditory signals in the inferior colliculus. Front. Neural Circuit. 6, 96. doi:10.3389/fncir.2012.00096.

Haegens, S., Golumbic, E.Z., 2018. Rhythmic facilitation of sensory processing: a critical review. Neurosci. Biobehav. Rev. 86, 150–165. doi:10.1016/j.neubiorev.2017.12.002.

Hamilton, L.S., Huth, A.G., 2020. The revolution will not be controlled: natural stimuli in speech neuroscience. Lang. Cogn. Neurosci. 35 (5), 573–582. doi:10.1080/23273798.2018.1499946.

Hasson, U., Egidi, G., Marelli, M., Willems, R.M., 2018. Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. Cognition 180, 135–157. doi:10.1016/j.cognition.2018.06.018.

Hebb, A.O., Ojemann, G.A., 2013. The thalamus and language revisited. Brain. Lang. 126 (1), 99–108. doi:10.1016/j.bandl.2012.06.010.

Hertrich, I., Dietrich, S., Ackermann, H., 2020. The Margins of the Language Network in the Brain [Review]. Front. Commun. 5 (93). doi:10.3389/fcomm.2020.519955.

Heyes, C., Catmur, C., 2022. What happened to mirror neurons? Perspect. Psychol. Sci. 17 (1), 1745691621990638. doi:10.1177/1745691621990638, Artn.

Hickok, G., Costanzo, M., Capasso, R., Miceli, G., 2011. The role of Broca's area in speech perception: Evidence from aphasia revisited. Brain Lang. 119 (3), 214–220. doi:10.1016/j.bandl.2011.08.001.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92 (1-2), 67–99. doi:10.1016/j.cognition.2003.10.011.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8 (5), 393–402. doi:10.1038/nrn2113.

Hickok, G., Rogalsky, C., Matchin, W., Basilakos, A., Cai, J., Pillay, S., Ferrill, M., Mickelsen, S., Anderson, S.W., Love, T., Binder, J., Fridriksson, J., 2018. Neural networks supporting audiovisual integration for speech: a large-scale lesion study. Cortex 103, 360–371. doi:10.1016/j.cortex.2018.03.030.

Hoffman, P., Pobric, G., Drakesmith, M., Ralph, M.A.L., 2012. Posterior middle temporal gyrus is involved in verbal and non-verbal semantic cognition: evidence from rTMS. Aphasiology 26 (9), 1119–1130. doi:10.1080/02687038.2011.608838.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532 (7600), 453–458. doi:10.1038/nature17637.

Iacoboni, M., 2008. The role of premotor cortex in speech perception: evidence from fMRI and rTMS. J. Physiol. Paris 102 (1-3), 31–34. doi:10.1016/j.jphysparis.2008.03.003.

James, T.W., Stevenson, R.A., Kim, S.Stain, B.E. (Ed.), 2012. Inverse effectiveness and BOLD fMRI. Ed. The New Handbook of Multisensory Processing.

Jung-Beeman, M., 2005. Bilateral brain processes for comprehending natural language. Trends Cogn. Sci. 9 (11), 512–518. doi:10.1016/j.tics.2005.09.009.

Komura, Y., Tamura, R., Uwano, T., Nishijo, H., Ono, T., 2005. Auditory thalamus integrates visual inputs into behavioral gains. Nat. Neurosci. 8 (9), 1203–1209. doi:10.1038/nn1528.

Kucera, H., Francis, W.N., 1967. Computational Analysis of Present-day American English. Brown University Press.

Lakatos, P., Gross, J., Thut, G., 2019. A new unifying account of the roles of neuronal entrainment. Curr. Biol. 29 (18), R890–R905. doi:10.1016/j.cub.2019.07.075.

Lakatos, P., Karmos, G., Mehta, A.D., Ulbert, I., Schroeder, C.E., 2008. Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320 (5872), 110–113. doi:10.1126/science.1154735.

Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U., 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. 31 (8), 2906–2915. doi:10.1523/JNEUROSCI.3684-10.2011.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech-perception revised. Cognition 21 (1), 1–36. doi:10.1016/0010-0277(85)90021-6.

Liuzzi, A.G., Aglinskas, A., Fairhall, S.L., 2020. General and feature-based semantic representations in the semantic network. Sci. Rep. 10 (1), 8931. doi:10.1038/s41598-020-65906-0, ARTN.

Luo, H., Liu, Z., Poeppel, D., 2010. Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biol. 8 (8), e1000445. doi:10.1371/journal.pbio.1000445.

Ma, W.J., Zhou, X., Ross, L.A., Foxe, J.J., Parra, L.C., 2009. Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. PLoS One 4 (3), e4638. doi:10.1371/journal.pone.0004638.

Macaluso, E., George, N., Dolan, R., Spence, C., Driver, J., 2004. Spatial and temporal factors during processing of audiovisual speech: a PET study. Neuroimage 21 (2), 725–732. doi:10.1016/j.neuroimage.2003.09.049.

MacLeod, A., Summerfield, Q., 1987. Quantifying the contribution of vision to speech perception in noise. Br. J. Audiol. 21 (2), 131–141. http://www.ncbi.nlm.nih.gov/pubmed/3594015.

Marchant, J.L., Ruff, C.C., Driver, J., 2012. Audiovisual synchrony enhances BOLD responses in a brain network including multisensory STS while also enhancing target-detection performance for both modalities. Hum. Brain Mapp. 33 (5), 1212–1224. doi:10.1002/hbm.21278.

Maus, B., van Breukelen, G.J.P., Goebel, R., Berger, M.P.F., 2010. Optimization of blocked designs in fMRI studies. Psychometrika 75 (2), 373–390. doi:10.1007/s11336-010-9159-3.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264 (5588), 746–748. http://www.ncbi.nlm.nih.gov/pubmed/1012311.

Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., Iacoboni, M., 2007. The essential role of premotor cortex in speech perception. Curr. Biol. 17 (19), 1692–1696. doi:10.1016/j.cub.2007.08.064.

Mercier, M.R., Foxe, J.J., Fiebelkorn, I.C., Butler, J.S., Schwartz, T.H., Molholm, S., 2013. Auditory-driven phase reset in visual cortex: human electrocorticography reveals mechanisms of early multisensory integration. Neuroimage 79, 19–29. doi:10.1016/j.neuroimage.2013.04.060.

Mercier, M.R., Molholm, S., Fiebelkorn, I.C., Butler, J.S., Schwartz, T.H., Foxe, J.J., 2015. Neuro-oscillatory phase alignment drives speeded multisensory response times: an electro-corticographic investigation. J. Neurosci. 35 (22), 8546–8557. doi:10.1523/JNEUROSCI.4527-14.2015.

Meredith, M.A., Stein, B.E., 1986. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. J. Neurophysiol. 56 (3), 640–662. doi:10.1152/jn.1986.56.3.640.

Miller, L.M., D'Esposito, M, 2005. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. J. Neurosci. 25 (25), 5884–5893. doi:10.1523/JNEUROSCI.0896-05.2005.

Molholm, S., Murphy, J.W., Bates, J., Ridgway, E.M., Foxe, J.J., 2020. Multisensory audiovisual processing in children with a sensory processing disorder (I): behavioral and electrophysiological indices under speeded response conditions. Front Integr Neurosci 14, 4. doi:10.3389/fnint.2020.00004.

Molholm, S., Ritter, W., Javitt, D.C., Foxe, J.J., 2004. Multisensory visual-auditory object recognition in humans: a high-density electrical mapping study. Cereb. Cortex 14 (4), 452–465. http://www.ncbi.nlm.nih.gov/pubmed/15028649; http://cercor.oxfordjournals.org/content/14/4/452.full.pdf.

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. Brain Res. Cogn. Brain Res. 14 (1), 115–128. http://www.ncbi.nlm.nih.gov/pubmed/12063135; http://ac.els-cdn.com/S0926641002000666/1-s2.0-S0926641002000666-main.pdf?_tid=d476e3d4-d24d-11e4-bd64-00000aab0f6b&acdnat=1427219424_49f04db69a10c36495d4b96f5036b7f7.

Munhall, K.G., Buchan, J.N., 2004. Something in the way she moves. Trends Cogn. Sci. 8 (2), 51–53. doi:10.1016/j.tics.2003.12.009.

Munhall, K.G., Gribble, P., Sacco, L., Ward, M., 1996. Temporal constraints on the McGurk effect. Percept. Psychophys. 58 (3), 351–362. doi:10.3758/bf03206811.

Munhall, K.G., Johnson, E.K., 2012. Speech perception: when to put your money where the mouth is. Curr. Biol. 22 (6), R190–R192. doi:10.1016/j.cub.2012.02.026.

Murase, M., Saito, D.N., Kochiyama, T., Tanabe, H.C., Tanaka, S., Harada, T., Aramaki, Y., Honda, M., Sadato, N., 2008. Cross-modal integration during vowel identification in audiovisual speech: a functional magnetic resonance imaging study. Neurosci. Lett. 434 (1), 71–76. doi:10.1016/j.neulet.2008.01.044.

Musacchia, G., Sams, M., Nicol, T., Kraus, N., 2006. Seeing speech affects acoustic information processing in the human brainstem. Exp. Brain Res. 168 (1-2), 1–10. doi:10.1007/s00221-005-0071-5.

Nath, A.R., Beauchamp, M.S., 2011. Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. J. Neurosci. 31 (5), 1704–1714. doi:10.1523/JNEUROSCI.4853-10.2011.

Navarra, J.Y., H, H., Werker, J.F., Soto-Faraco, Salvador, 2012. Multisensory interactions in speech perception. The New Handbook of Multisensory Processing. MIT Press, pp. 435–452.

Nichols, T., Brett, M., Andersson, J., Wager, T., Poline, J.B., 2005. Valid conjunction inference with the minimum statistic. Neuroimage 25 (3), 653–660. doi:10.1016/j.neuroimage.2004.12.005.

Noesselt, T., Bergmann, D., Heinze, H.J., Munte, T., Spence, C., 2012. Coding of multisensory temporal patterns in human superior temporal sulcus. Front. Integr. Neurosci. 6, 64. doi:10.3389/fnint.2012.00064.

Noesselt, T., Rieger, J.W., Schoenfeld, M.A., Kanowski, M., Hinrichs, H., Heinze, H.J., Driver, J., 2007. Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. J. Neurosci. 27 (42), 11431–11441. doi:10.1523/JNEUROSCI.2252-07.2007.

Noesselt, T., Tyll, S., Boehler, C.N., Budinger, E., Heinze, H.J., 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 (41), 13609–13623. doi:10.1523/JNEUROSCI.4524-09.2010.

Nozawa, G., Reuter-Lorenz, P.A., Hughes, H.C., 1994. Parallel and serial processes in the human oculomotor system: bimodal integration and express saccades. Biol. Cybern. 72 (1), 19–34. http://www.ncbi.nlm.nih.gov/pubmed/7880912.

Nuttall, H.E., Kennedy-Higgins, D., Devlin, J.T., Adank, P., 2017. The role of hearing ability and speech distortion in the facilitation of articulatory motor cortex. Neuropsychologia 94, 13–22. doi:10.1016/j.neuropsychologia.2016.11.016.

Nuttall, H.E., Kennedy-Higgins, D., Devlin, J.T., Adank, P., 2018. Modulation of intra- and inter-hemispheric connectivity between primary and premotor cortex during speech perception. Brain Lang. 187, 74–82. doi:10.1016/j.bandl.2017.12.002.

Ojanen, V., Mottonen, R., Pekkola, J., Jaaskelainen, I.P., Joensuu, R., Autti, T., Sams, M., 2005. Processing of audiovisual speech in Broca's area. Neuroimage 25 (2), 333–338. doi:10.1016/j.neuroimage.2004.12.001.

Okada, K., Venezia, J.H., Matchin, W., Saberi, K., Hickok, G., 2013. An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. PLoS One 8 (6), e68959. doi:10.1371/journal.pone.0068959.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9 (1), 97–113. http://www.ncbi.nlm.nih.gov/pubmed/5146491.

Pearson, J., 2019. The human imagination: the cognitive neuroscience of visual mental imagery. Nat. Rev. Neurosci. 20 (10), 624–634. doi:10.1038/s41583-019-0202-9.

Peelle, J., Grossman, M., 2008. Language processing in frontotemporal dementia: a brief review. Lang. Linguistic. Compass 2, 18–35. doi:10.1111/j.1749-818X.2007.00047.x.

Peelle, J.E.Katz, W.F., Assmann, P.F. (Eds.), 2019. The neural basis for auditory and audiovisual speech perception. Ed. The Routledge Handbook of Phonetics.

Peelle, J.E., Spehar, B., Jones, M.S., McConkey, S., Myerson, J., Hale, S., Sommers, M.S., Tye-Murray, N., 2021. Increased connectivity among sensory and motor regions during visual and audiovisual speech perception. J. Neurosci. doi:10.1523/JNEUROSCI.0114-21.2021.

Price, C.J., 2010. The anatomy of language: a review of 100 fMRI studies published in 2009. Ann. N Y Acad. Sci. 1191, 62–88. doi:10.1111/j.1749-6632.2010.05444.x.

Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage 62 (2), 816–847. doi:10.1016/j.neuroimage.2012.04.062.

Puce, A., Allison, T., Bentin, S., Gore, J.C., McCarthy, G., 1998. Temporal cortex activation in humans viewing eye and mouth movements. J. Neurosci. 18 (6), 2188–2199. https://www.ncbi.nlm.nih.gov/pubmed/9482803.

Puce, A., Syngeniotis, A., Thompson, J.C., Abbott, D.F., Wheaton, K.J., Castiello, U., 2003. The human temporal lobe integrates facial form and motion: evidence from fMRI and ERP studies. Neuroimage 19 (3), 861–869. doi:10.1016/s1053-8119(03)00189-7.

Pulvermuller, F., Fadiga, L., 2010. Active perception: sensorimotor circuits as a cortical basis for language. Nat. Rev. Neurosci. 11 (5), 351–360. doi:10.1038/nrn2811.

Pulvermuller, F., Huss, M., Kherif, F., Martin, F.M.D.P., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. U.S.A. 103 (20), 7865–7870. doi:10.1073/pnas.0509989103.

Rauschecker, J.P., 2012. Ventral and dorsal streams in the evolution of speech and language. Front. Evol. Neurosci. 4, 7. doi:10.3389/fnevo.2012.00007.

Reale, R.A., Calvert, G.A., Thesen, T., Jenison, R.L., Kawasaki, H., Oya, H., Howard, M.A., Brugge, J.F., 2007. Auditory-visual processing represented in the human superior temporal gyrus. Neuroscience 145 (1), 162–184. doi:10.1016/j.neuroscience.2006.11.036.

Richie, C., Kewley-Port, D., 2008. The effects of auditory-visual vowel identification training on speech recognition under difficult listening conditions. J. Speech Lang. Hear. Res. 51 (6), 1607–1619. doi:10.1044/1092-4388(2008/07-0069).

Rizzolatti, G., Arbib, M.A., 1998. Language within our grasp. Trends Neurosci. 21 (5), 188–194. doi:10.1016/S0166-2236(98)01260-0.

Rogalsky, C., Hickok, G., 2011. The role of Broca's area in sentence comprehension. J. Cogn. Neurosci. 23 (7), 1664–1680. doi:10.1162/jocn.2010.21530.

Ross, L.A., Del Bene, V.A., Molholm, S., Frey, H.P., Foxe, J.J., 2015. Sex differences in multisensory speech processing in both typically developing children and those on the autism spectrum. Front. Neurosci. 9, 185. doi:10.3389/fnins.2015.00185.

Ross, L.A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., Foxe, J.J., 2011. The development of multisensory speech perception continues into the late childhood years. Eur. J. Neurosci. 33 (12), 2329–2337. doi:10.1111/j.1460-9568.2011.07685.x.

Ross, L.A., Saint-Amour, D., Leavitt, V.M., Javitt, D.C., Foxe, J.J., 2007. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb. Cortex 17 (5), 1147–1153. doi:10.1093/cercor/bhl024.

Ross, L.A., Saint-Amour, D., Leavitt, V.M., Molholm, S., Javitt, D.C., Foxe, J.J., 2007. Impaired multisensory processing in schizophrenia: deficits in the visual enhancement of speech comprehension under noisy environmental conditions. Schizophr. Res. 97 (1-3), 173–183. doi:10.1016/j.schres.2007.08.008.

Rowland, B.A., Quessy, S., Stanford, T.R., Stein, B.E., 2007. Multisensory integration shortens physiological response latencies. J. Neurosci. 27 (22), 5879–5884. doi:10.1523/JNEUROSCI.4986-06.2007.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., Foxe, J.J., 2007. Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. Neuropsychologia 45 (3), 587–597. doi:10.1016/j.neuropsychologia.2006.03.036.

Schomers, M.R., Pulvermuller, F., 2016. Is the sensorimotor cortex relevant for speech perception and understanding? An integrative review. Front. Hum. Neurosci. 10, 435. doi:10.3389/fnhum.2016.00435.

Schroeder, C.E., Lakatos, P., 2009. Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. 32 (1), 9–18. doi:10.1016/j.tins.2008.09.012.

Scott, S.K., McGettigan, C., Eisner, F., 2009. A little more conversation, a little less action–candidate roles for the motor cortex in speech perception. Nat. Rev. Neurosci. 10 (4), 295–302. doi:10.1038/nrn2603.

Sekiyama, K., Kanno, I., Miura, S., Sugita, Y., 2003. Auditory-visual speech perception examined by fMRI and PET. Neurosci. Res. 47 (3), 277–287. http://www.ncbi.nlm.nih.gov/pubmed/14568109; http://ac.els-cdn.com/S0168010203002141/1-s2.0-S0168010203002141-main.pdf?_tid=b93bc77c-db06-11e3-8a59-00000aab0f27&acdnat=1400030947_9d708076e7cdc88b0c032fd02d51e07e; http://ac.els-cdn.com/S0168010203002141/1-s2.0-S0168010203002141-main.pdf?_tid=afef6ef2-1c16-11e4-a960-00000aacb35e&acdnat=1407184629_66c42f8746b68d9cfc2a2e44dce8caae.

Senkowski, D., Saint-Amour, D., Gruber, T., Foxe, J.J., 2008. Look who's talking: the deployment of visuo-spatial attention during multisensory speech processing under noisy environmental conditions. Neuroimage 43 (2), 379–387. doi:10.1016/j.neuroimage.2008.06.046.

Skipper, J.I., Nusbaum, H.C., Small, S.L., 2005. Listening to talking faces: motor cortical activation during speech perception. Neuroimage 25 (1), 76–89. doi:10.1016/j.neuroimage.2004.11.006.

Smith, S., Jenkinson, M., Beckmann, C., Miller, K., Woolrich, M., 2007. Meaningful design and contrast estimability in FMRI. Neuroimage 34 (1), 127–136. doi:10.1016/j.neuroimage.2006.09.019.

Sperdin, H.F., Cappe, C., Foxe, J.J., Murray, M.M., 2009. Early, low-level auditory-somatosensory multisensory interactions impact reaction time speed. Front. Integr. Neurosci. 3, 2. doi:10.3389/neuro.07.002.2009.

Stasenko, A., Bonn, C., Teghipco, A., Garcea, F.E., Sweet, C., Dombovy, M., McDonough, J., Mahon, B.Z., 2015. A causal test of the motor theory of speech perception: a case of impaired speech production and spared speech perception. Cognit. Neuropsychol. 32 (2), 38–57. doi:10.1080/02643294.2015.1035702.

Stein, B.E., Huneycutt, W.S., Meredith, M.A., 1988. Neurons and behavior: the same rules of multisensory integration apply. Brain Res. 448 (2), 355–358. http://www.ncbi.nlm.nih.gov/pubmed/3378157; http://ac.els-cdn.com/0006899388912760/1-s2.0-0006899388912760-main.pdf?_tid=6a19fa6e-dc65-11e3-889f-00000aacb360&acdnat=1400181568_c6d2e7d9e86dbb053facdaefa5ba3508.

Stein, B.E., Meredith, M.A., 1993. The Merging of the Senses. MIT Press.

Stein, B.E., Meredith, M.A., Huneycutt, W.S., McDade, L., 1989. Behavioral indices of multisensory integration: orientation to visual cues is affected by auditory stimuli. J. Cogn. Neurosci. 1 (1), 12–24. doi:10.1162/jocn.1989.1.1.12.

Stein, B.E., Meredith, M.A., Wallace, M.T., 1993. The visually responsive neuron and beyond: multisensory integration in cat and monkey. Prog. Brain Res. 95, 79–90. http://www.ncbi.nlm.nih.gov/pubmed/8493355.

Stein, B.E., Stanford, T.R., 2008. Multisensory integration: current issues from the perspective of the single neuron. Nat. Rev. Neurosci. 9 (4), 255–266. doi:10.1038/nrn2331.

Stevenson, R.A., Altieri, N.A., Kim, S., Pisoni, D.B., James, T.W., 2010. Neural processing of asynchronous audiovisual speech perception. Neuroimage 49 (4), 3308–3318. doi:10.1016/j.neuroimage.2009.12.001.

Stevenson, R.A., Bushmakin, M., Kim, S., Wallace, M.T., Puce, A., James, T.W., 2012. Inverse effectiveness and multisensory interactions in visual event-related potentials with audiovisual speech. Brain Topogr. 25 (3), 308–326. doi:10.1007/s10548-012-0220-7.

Stevenson, R.A., James, T.W., 2009. Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. Neuroimage 44 (3), 1210–1223. doi:10.1016/j.neuroimage.2008.09.034.

Stevenson, R.A., Kim, S., James, T.W., 2009. An additive-factors design to disambiguate neuronal and areal convergence: measuring multisensory interactions between audio, visual, and haptic sensory streams using fMRI. Exp. Brain Res. 198 (2-3), 183–194. doi:10.1007/s00221-009-1783-8.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.

Tjan, B.S., Chao, E., Bernstein, L.E., 2014. A visual or tactile signal makes auditory speech detection more efficient by reducing uncertainty. Eur. J. Neurosci. 39 (8), 1323–1331. doi:10.1111/ejn.12471.

Turken, A.U., Dronkers, N.F., 2011. The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. Front. Syst. Neurosci. 5. doi:10.3389/fnsys.2011.00001, ARTN 1.

Tye-Murray, N., Sommers, M.S., Spehar, B., 2007. The effects of age and gender on lipreading abilities. J. Am. Acad. Audiol. 18 (10), 883–892. doi:10.3766/jaaa.18.10.7.

van Atteveldt, N.M., Formisano, E., Blomert, L., Goebel, R., 2007. The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. Cereb. Cortex 17 (4), 962–974. doi:10.1093/cercor/bhl007.

van de Rijt, L.P.H., Roye, A., Mylanus, E.A.M., van Opstal, A.J., van Wanrooij, M.M., 2019. The principle of inverse effectiveness in audiovisual speech perception. Front. Hum. Neurosci. 13, 335. doi:10.3389/fnhum.2019.00335.

Van der Burg, E., Olivers, C.N., Bronkhorst, A.W., Theeuwes, J., 2008. Pip and pop: nonspatial auditory signals improve spatial visual search. J. Exp. Psychol. Hum. Percept. Perform. 34 (5), 1053–1065. doi:10.1037/0096-1523.34.5.1053.

Van Engen, K.J., Peelle, J.E., 2014. Listening effort and accented speech. Front. Hum. Neurosci. 8, 577. doi:10.3389/fnhum.2014.00577.

van Wassenhove, V., Grant, K.W., Poeppel, D., 2007. Temporal window of integration in auditory-visual speech perception. Neuropsychologia 45 (3), 598–607. doi:10.1016/j.neuropsychologia.2006.01.001.

Vanni, S., Tanskanen, T., Seppa, M., Uutela, K., Hari, R., 2001. Coinciding early activation of the human primary visual cortex and anteromedial cuneus. Proc. Natl. Acad. Sci. U.S.A. 98 (5), 2776–2780. doi:10.1073/pnas.041600898.

Wallace, M.T., Meredith, M.A., Stein, B.E., 1998. Multisensory integration in the superior colliculus of the alert cat. J. Neurophysiol. 80 (2), 1006–1010. http://www.ncbi.nlm.nih.gov/pubmed/9705489.

Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nat. Neurosci. 7 (7), 701–702. doi:10.1038/nn1263.

Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J., McCarthy, G., 2003. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb. Cortex 13 (10), 1034–1043. http://www.ncbi.nlm.nih.gov/pubmed/12967920; http://cercor.oxfordjournals.org/content/13/10/1034.full.pdf.

Xu, J., Kemeny, S., Park, G., Frattali, C., Braun, A., 2005. Language in context: emergent features of word, sentence, and narrative comprehension. Neuroimage 25 (3), 1002–1015. doi:10.1016/j.neuroimage.2004.12.013.

Xu, J., Yu, L., Rowland, B.A., Stanford, T.R., Stein, B.E., 2014. Noise-rearing disrupts the maturation of multisensory integration. Eur. J. Neurosci. 39 (4), 602–613. doi:10.1111/ejn.12423.

Xu, Y., He, Y., Bi, Y., 2017. A Tri-network model of human semantic processing. Front. Psychol. 8, 1538. doi:10.3389/fpsyg.2017.01538.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods 8 (8), 665–670. doi:10.1038/nmeth.1635.

Yu, L., Xu, J., Rowland, B.A., Stein, B.E., 2013. Development of cortical influences on superior colliculus multisensory neurons: effects of dark-rearing. Eur. J. Neurosci. 37 (10), 1594–1601. doi:10.1111/ejn.12182.

Zion Golumbic, E., Cogan, G.B., Schroeder, C.E., Poeppel, D, 2013. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". J. Neurosci. 33 (4), 1417–1426. doi:10.1523/JNEUROSCI.3675-12.2013.

Zoefel, B., ten Oever, S., Sack, A.T., 2018. The involvement of endogenous neural oscillations in the processing of rhythmic input: more than a regular repetition of evoked neural responses. Front. Neurosci. 12. doi:10.3389/fnins.2018.00095, ARTN 95.