Published in final edited form as:
Neuroimage. 2010 February 15; 49(4): 3308. doi:10.1016/j.neuroimage.2009.12.001.

Neural processing of asynchronous audiovisual speech perception

Ryan A. Stevenson[1,2], **Nicholas A. Altieri**[1], **Sunah Kim**[2,3], **David B. Pisoni**[1,3], and **Thomas W. James**[1,2,3]

[1] Department of Psychological and Brain Sciences, Indiana University

[2] Program in Neuroscience, Indiana University

[3] Cognitive Science Program, Indiana University

Abstract

The temporal synchrony of auditory and visual signals is known to affect the perception of an external event, yet it is unclear what neural mechanisms underlie the influence of temporal synchrony on perception. Using parametrically varied levels of stimulus asynchrony in combinations with BOLD fMRI, we identified two anatomically distinct subregions of multisensory superior temporal cortex (mSTC) that showed qualitatively distinct BOLD activation patterns. A synchrony-defined subregion of mSTC (synchronous > asynchronous) responded only when auditory and visual stimuli were synchronous, whereas a bimodal subregion of mSTC (auditory > baseline and visual > baseline) showed significant activation to all presentations, but showed monotonically increasing activation with increasing levels of asynchrony. The presence of two distinct activation patterns suggests that the two subregions of mSTC may rely on different neural mechanisms to integrate audiovisual sensory signals. An additional whole-brain analysis revealed a network of regions responding more with synchronous than asynchronous speech, including right mSTC, and bilateral superior colliculus, fusiform gyrus, lateral occipital cortex, and extrastriate visual cortex. The spatial location of individual mSTC ROIs was much more variable in the left than right hemisphere, suggesting that individual differences may contribute to the right lateralization of mSTC in a group SPM. These findings suggest that bilateral mSTC is composed of distinct multisensory subregions that integrate audiovisual speech signals through qualitatively different mechanisms, and may be differentially sensitive to stimulus properties including, but not limited to, temporal synchrony.

Keywords

Multisensory; Integration; STS; fMRI; synchrony; superior colliculus

Introduction

The ability of an individual to integrate multiple sensory signals generated from a single external source depends on the properties of those sensory signals. Perception of audiovisual

Correspondence to: Ryan A. Stevenson Department of Psychological and Brain Sciences Indiana University 1101 East Tenth Street, Room 293 Bloomington, IN 47405 stevenra@indiana.edu Fax: (812) 855-4691 Phone: (812) 856-1926.

signals is strongly influenced by the relative timing, or the temporal synchrony, of auditory and visual signals (Miller and D'Esposito, 2005; van Atteveldt et al., 2007; van Wassenhove et al., 2007). Manipulations of temporal synchrony also affect neural responses to multisensory audiovisual stimuli (King and Palmer, 1985; Macaluso et al., 2004; Meredith et al., 1987; Miller and D'Esposito, 2005; Stein et al., 1993). One compelling example of the role that temporal synchrony plays in multisensory speech perception can be seen in the case study of AWF, a stroke victim who reported an impairment of his speech perception. When watching someone talk, AWF perceives the auditory and visual components of the talker's utterance as separate, temporally asynchronous events (Hamilton et al., 2006). His impairment interferes with his ability to understand spoken language: speech intelligibility is worse when AWF can see a talker's lip movements than when he hears the sounds alone. The impact of visible lip movements on AWF's ability to perceive and integrate multisensory speech signals is quite different from healthy listeners and hearing-impaired listeners and provides an example of the importance of temporal processing in multisensory perception.

For healthy individuals, visual information provided by a talker's face through lip-reading usually enhances the intelligibility of auditory speech. It is now well established that the addition of semantically-congruent visual information to the auditory speech signal significantly increases accuracy scores (audiovisual enhancement) across a considerable range of signal-to-noise ratios (Grant et al., 1998; Sumby and Pollack, 1954; Summerfield, 1987), and facilitates neural processing in the auditory cortex (van Wassenhove et al., 2005). Also, this interaction goes beyond mutual facilitation across sensory streams; it produces perceptual fusion (also known as binding), where the auditory and visual sensory streams are combined to produce a single percept of an external event, in this case the speaker's utterance. Thus, in natural conversations between healthy individuals, the subjective experience of speech is that of an integrated, unified percept (Gaver, 1993).

As the example of AWF shows, the temporal synchrony of auditory and visual speech signals plays an important role in this perceptual fusion of speech signals. This role can be clearly seen through the impact that temporal synchrony has on the McGurk effect. The McGurk effect is an audiovisual speech illusion in which the combination of incongruent but temporally-synchronous visual and auditory speech signals produces a percept not represented by either unisensory speech signal (Green and Kuhl, 1989; McGurk and MacDonald, 1976). For instance, the presentation of an auditory /b/ (bilabial stop consonant) combined with a visual /g/ (velar stop consonant) usually yields the fused percept /d/ – a percept not specified in either the auditory or visual signal. This type of perceptual fusion, however, is greatly influenced by the temporal synchronization of the auditory and visual speech signals. When the auditory and visual signals become more asynchronous, perceptual fusion is less likely to occur (van Atteveldt et al., 2007; van Wassenhove et al., 2007), although in some cases perceptual fusion may occur in the absence of perceived synchrony (Soto-Faraco and Alsius, 2009). Based on these results in healthy listeners, it is possible that patients with impaired perceptual fusion deficits such as AWF may represent a specific problem with temporal processing of the input signals.

For optimal perceptual fusion to occur in healthy listeners, auditory and visual speech signals must be synchronized within a relatively narrow time window. Experimental manipulation of temporal synchrony with speech stimuli does not disrupt audiovisual fusion or perceived synchrony if the asynchrony is within a certain tolerance (150 to 450 ms, depending on experimental paradigm and stimulus; Conrey and Pisoni, 2006; Conrey and Pisoni, 2004; Steinmetz, 1996; van Wassenhove et al., 2007). That is, if the offset of the auditory and visual signals is within a limited range, the two sensory inputs are often still perceived to be synchronous even though they are not. The time window in which asynchronous speech is perceived as synchronous is often found to be longer when the visual input precedes the

auditory input than when the auditory input precedes the visual (Conrey and Pisoni, 2006; Conrey and Pisoni, 2004; Grant et al., 2004; Miller and D'Esposito, 2005; van Wassenhove et al., 2007).

While behavioral measurements of perceptual fusion remain unaffected by synchrony manipulations within a considerable range, the neural mechanisms that underlie this phenomenon in humans remain unknown. In the case of AWF, while intriguing, there was no conclusive evidence of a neural mechanism or substrate responsible for his impairment. An MRI performed on AWF was unremarkable, and a PET scan reported only parietal hypoperfusion, which suggests that the impairment was due to a possible vascular incident such as a stroke in his parietal lobe. Previous findings from neural recordings in non-human animals do, however, provide some insights into the influence of temporal synchrony on multisensory neurons. These non-human animal studies suggest the existence of a network of neurons whose neuronal firing rates are modulated by changes in temporal synchrony, even when those changes in synchrony are within the time window of perceptual fusion (King and Palmer, 1985; Meredith et al., 1987; Stein et al., 1993). Research on humans, performed with PET and fMRI measurements, provide additional evidence for a network of brain regions that process multisensory speech, but research on the influence of temporal-synchrony on those networks is limited and has yielded conflicting results (Macaluso et al., 2004; Miller and D'Esposito, 2005).

For instance, in a PET study, Macaluso and colleagues (2004) identified a network of brain regions that responded preferentially to synchronous speech over asynchronous (audio preceding visual by 240 ms) speech including bilateral fusiform gyrus (FG), right medial lingual gyrus, left STS, bilateral lateral occipital complex (LO), and bilateral dorsal occipital cortex. Miller and colleagues (2005), on the other hand, compared the blood-oxygen-level-dependant (BOLD) fMRI activations after presenting participants with synchronous and asynchronous speech (both audio- and visual-first presentations at each individual's 50% detection threshold, mean audio lead = 141 ms, mean visual lead = 215 ms) and failed to identify any brain regions that responded more to the synchronous condition. Conversely, they reported a network of brain regions that responded more to the asynchronous conditions, including the superior colliculus (SC), anterior insula, and anterior intraparietal sulcus (IPS).

It is important to note that the majority of the regions Macaluso et al. (2004) and Miller and D'Espisito (2005) identified have been previously implicated in multisensory convergence. One of these regions producing conflicting results is the multisensory superior temporal cortex (mSTC), including perhaps the most often studied of these multisensory regions, STS. One possible explanation for the disparate findings in the two studies described above is that mSTC is not a single functional region with uniform activation patterns. Indeed, multisensory STC is comprised of anatomical subregions that respond differentially with certain classes of stimuli (Puce et al., 1998; Scott et al., 2000; Stevenson and James, 2009), and additionally has been described as being composed of patches of subregions that process unisensory inputs (including auditory, visual, and somatosensory inputs) that feed into multisensory subregions (Beauchamp et al., 2004a). Given these previous findings suggesting nonuniformity within mSTC, as well as the anatomical difference between STS regions in the two previously described synchrony studies (x, y, and, z Talairach coordinates were −46, −28, 0, in Miller et al. (2005), and −64, −58, 0, in Macaluso et al. (2004)), it is possible that the different results found with variations in temporal synchrony may reflect activation from different subregions within mSTC.

In the present study, we investigated the neural substrates involved in processing temporal synchrony and asynchrony with audiovisual speech signals. First, we used audiovisual speech stimuli presented at parametrically varied temporal alignments to identify a network of brain

regions in which the activation level reflected the level of temporal alignment of the auditory and visual signals. Regions include the right mSTC and FG, as well as bilateral SC, LO, and earlier visual areas. We identified two distinct activation profiles. The first profile showed a continuous change in activation with increases in asynchrony, and may represent brain regions modulated by either temporal asynchrony or the highly correlated rate of perceptual fusion. The second profile revealed a discrete change in activation with physically synchronous stimuli, and may represent brain regions involved in synchrony detection.

Second, we identified anatomically and functionally distinct regions within mSTC and examined their activation profiles with changes in temporal asynchrony. Bimodal mSTC, which was defined based on significant activation with auditory *and* visual unisensory stimuli, responded in a graded fashion to the parametrically-varied levels of stimulus asynchrony (i.e., activation parametrically increased as the level of asynchrony increased). Synchrony mSTC, which was identified by an interaction across synchrony levels, responded like a synchrony detector, showing activation only when the physical stimulus was temporally synchronous, regardless of perceived synchrony. The identification of two functionally and anatomically distinct regions within mSTC further suggests that mSTC is a heterogeneous group of functionally-specialized subregions. This functional dichotomy also may provide new insights into possible brain mechanisms underlying impairments in temporal processing similar to those reported for AWF's.

Methods and Materials

Procedure Overview

Participants took part in a two-phase experimental procedure. First, participants' perceived synchrony of audiovisual speech tokens of isolated words were measured, with the synchronization of the auditory and visual components parametrically manipulated from 400 ms with video preceding audio (V-A) to 400 ms with audio preceding video (A-V) in 100 ms intervals. Next, participants' BOLD activation were measured in two paradigms; with the fast event-related presentation of the same single, word, audiovisual speech tokens with parametrically-varied synchronies (referred to as "experimental runs"), and also with blocked, visual and auditory unisensory speech presentations (referred to as "functional localizer runs").

Participants

Participants included 8 right-handed native English speakers (4 female, mean age = 24.1). Our experimental protocol was approved by the Indiana University Institutional Review Board and Human Subjects Committee.

Stimulus Materials

Stimuli included dynamic, audiovisual (AV) recordings of a female speaker saying ten highly familiar nouns (see Figure 1). Stimuli were selected from a previously published database, The Hoosier Audiovisual Multi-Talker Database (Sheffert et al., 1996). All stimuli were spoken by speaker F1. We selected words that were monosyllabic, had the highest levels of accuracy on both visual-only and audio-only recognition (Lachs and Hernandez, 1998), and resided in low-density lexical neighborhoods (Luce and Pisoni, 1998;Sheffert et al., 1996). From the set of words that matched these criteria, we selected 10 items that fell into two easily distinguishable semantic categories, and had approximately equal mean word lengths across categories. The two categories were chosen based on the above mentioned criteria, consisting of body parts (face, leg, mouth, neck, and teeth) and environmental words (beach, dirt, rain, rock, and soil). Mean body part word duration was 1.562 s, and mean environmental-word duration was 1.582 s. Audio signal levels were measured as root mean square contrasts and equated across tokens using MATLAB 5.2 (MATHWORKS Inc., Natick, MA).

All stimuli used in this study were presented using MATLAB 5.2 (MATHWORKS Inc., Natick, MA) software with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997), running on a Macintosh computer. Visual stimuli were projected onto a frosted glass screen using a Mitsubishi XL30U projector. Visual stimuli were 200 × 200 pixels and subtended 4.8 × 4.8° of visual angle. Audio stimuli were presented using pneumatic headphones.

To ensure the precision of audiovisual offsets, a simulation of 1000 trials were run at each offset level (a total of 9000 trials). Across stimulus offsets, 95% of all trials were within 10 ms of the predicted offset, and 98% within 20 ms. No two offset conditions overlapped in any of the 9000 trials. Precision at each offset level is shown in Figure 1e, with different colored bars representing each offset condition.

Behavioral pre-scan procedures

Participants' individual sensitivity to asynchrony was measured prior to scanning while in an MRI simulator designed to mimic the actual fMRI. Participants were presented with the audiovisual spoken-word tokens described above, with the temporal synchrony varied parametrically from 400 ms V-A to 400 ms A-V in 100 ms increments (see Figure 1b and c). This included a condition in which the onset asynchrony was 0 ms, which we will refer to as the synchronous condition (see Figure 1a). Participants performed a two-alternative forced-choice (2AFC) decision: judging whether the spoken word was synchronous or asynchronous. During this task, pre-recorded scanner noise was played at an equal signal level to the actual MRI. 30 trials were presented for each level of onset asynchrony, and responses were collected by a button press.

Scanning procedures

Each imaging session included two phases: functional localizer runs and experimental runs. Functional localizer runs consisted of stimuli presented in a blocked stimulus design while participants completed an orthogonal, 2AFC categorization task (body-part or environmental word). Each run began with the presentation of a fixation cross for 12 s followed by six blocks of audio, visual, or audiovisual stimuli. The auditory and visual components of the localizer stimuli were always semantically congruent, a feature that has been shown to be a critical factor in multisensory enhancement (Laurienti et al., 2004), as well as temporally synchronous. Each run included two 16 s blocks of each stimulus type, with blocks consisting of eight stimulus presentations, separated by 0.1 s inter-stimuli intervals (ISI). New blocks began every 28 s separated by fixation. Runs ended with 12 s of fixation. Block orders were counterbalanced across runs and participants. Each participant completed two functional localizer runs.

During experimental runs, stimuli were presented in a fast event-related design in which participants performed the same a 2AFC semantic categorization task as in the localizer runs. Each run included presentations of audiovisual stimuli at each level of onset asynchrony (recall that audiovisual stimuli ranged parametrically from 400 ms V-A to 400 ms A-V in 100 ms intervals). Runs began with the presentation of a fixation cross for 12 s, followed by seven trials at each asynchrony level, for a total of 63 trials per run. For the seven trials of each stimulus type, four trials were preceded by a two-second ISI, two trials preceded by a four-second ISI, and one trial by a six-second ISI, with ISIs consisting of a static visual fixation cross. Runs concluded with 12 s of fixation. Trial and ISI orders were counterbalanced across runs and run order was counterbalanced across participants. Each participant completed seven experimental runs, for a total of 49 trials per condition.

Due to the different temporal offsets, stimulus presentation times for asynchronous stimuli were increased by the exact amount of the offset (for example, a stimulus in which the audio

onset preceded the visual onset by 400 ms has a total presentation time that was 400 ms longer than the synchronous condition, compare Figure 1a to Figures 1b and c for a visual example). To ensure that any differences observed with onset asynchrony were not due to increased stimulus presentation time, a subset of the participants (N = 4, 2 female, mean age = 25.0) completed a second scanning session. This session included functional localizers identical to those described above, and an additional set of modified experimental runs. Instead of varying the asynchrony and stimulus length, audiovisual stimuli were presented synchronously with parametrically-varied presentation times ranging from the synchronous presentation time in the above described experiment to 400 ms longer presentation time in 100 ms increments (see Figure 1d). The variation of presentation length in this control experiment exactly matched the presentation times in the first experiment, with no variation in synchrony, allowing us to infer that any effect found in the first experiment were in fact due to variations in synchrony and rule out that any effect observed in the first experiment were due to increased presentation lengths caused by the temporal offset. Presentations were lengthened by slowing the screen refresh rate for visual manipulation and using a slower bit-per-second presentation rate for the audio manipulation. As such, these presentations matched the experimental runs in total presentation time, but were always synchronous (compare Figures 1b and c to Figure 1d for a visual example).

Imaging parameters and analysis

Imaging was carried out using a Siemens Magnetom Trio 3-T whole body scanner, and collected on an eight-channel phased-array head coil. The field of view was 22 × 22 × 9.9 cm, with an in plane resolution of 64 × 64 pixels and 33 axial slices per volume (whole brain), creating a voxel size of 3.44 × 3.44 × 3.4 mm, re-sampled at 3 × 3 × 3 mm. Images were collected using a gradient echo EPI (TE = 30 ms, TR = 2000 ms, flip angle = 70°) for BOLD imaging. High-resolution T1-weighted anatomical volumes were acquired using Turbo-flash 3-D (TI = 1,100 ms, TE = 3.93 ms, TR = 14.375 ms, Flip Angle = 12°) with 160 sagittal slices with a thickness of 1 mm and field of view of 224 × 256 (voxel size = 1 × 1 × 1 mm).

Imaging data were pre-processed using Brain Voyager ™ 3-D analysis tools. Functional data underwent a linear trend removal, 3-D spatial Gaussian filtering (FWHM 6 mm), slice scan time correction, and 3-D motion correction. Anatomical volumes were transformed into a common stereotactic space (Talaraich and Tournoux, 1988) using an 8-point affine transformation. Functional data were aligned to the first volume of the run closest in time to the anatomical data collection. Each functional run was then aligned to the transformed anatomical volumes, transforming the functional data to a common stereotactic space across participants.

Whole-brain, random-effects (RFX) statistical parametric maps (SPM) were calculated using the Brain Voyager™ general linear model (GLM) procedure. The design matrix was assembled from separate predictors for each audiovisual presentation at each level of asynchrony (9 predictors total, 2 s events) modeled using a canonical hemodynamic response function (Glover, 1999). This was not a deconvolution design matrix. Event-related averages (ERA), consisting of aligning and averaging all trials from each condition to stimulus onset were created based on onset asynchrony for both the localizer and the experimental study. Hemodynamic BOLD activations were defined as the arithmetic mean of the time course within a time window 6–16 s after block onset for the localizer runs, and a window of 4-6 s after trial onset for the fast event-related experimental runs.

Results

Behavioral Results

Perception of audiovisual asynchrony was measured behaviorally for each individual prior to scanning in an fMRI simulator. As onset asynchronies approached 0 ms, participants were more likely to judge the stimulus as synchronous, with synchrony judgments decreasing asymmetrically as the onset asynchrony increased (see Figure 2a for values associated with each onset asynchrony). The drop in perceived synchrony was greater with V-A than with A-V stimulus presentations, replicating earlier studies of asynchrony detection (Conrey and Pisoni, 2006; Conrey and Pisoni, 2004; Dixon and Spitz, 1980; Grant and Greenberg, 2001; Grant et al., 2004; McGrath and Summerfield, 1985; Miller and D'Esposito, 2005; Munhall et al., 1996; van Wassenhove et al., 2007). Accuracies were also collected for the 2AFC semantic categorization task that participants completed during scanning. Mean accuracies were calculated for each individual participant with each temporal offset (mean = 94%, SD = 6%). No significant difference in accuracy of semantic categorization was found between any pairwise comparison of synchrony levels.

Individual analyses

To investigate multisensory activations, whole-brain fixed-effects (FFX) SPMs were calculated for each individual. Multisensory STC (mSTC) was defined bilaterally in each individual using two distinct functional definitions. The first manner in which mSTC was defined was through the use of a synchrony > asynchrony contrast (see Table 1 for average ROI locations). Synchronous trials were defined as those trials with an onset asynchrony of 0 ms, while asynchronous trials were defined as those trials with an onset asynchrony of 300 ms or greater in either direction, outside the perceived synchrony level in both A-V and V-A presentations with these stimuli, as measured in a previous experiment (Conrey and Pisoni, 2006;Conrey and Pisoni, 2004). These asynchronous levels with 300-400 ms have as such been labeled in red in Figures 3 and 4. All individual participants exhibited *bilateral* mSTC activations with significantly more activation with the synchronous than asynchronous trials, a subregion of mSTC that we will refer to as synchrony-defined mSTC (S-mSTC). ERAs were extracted from each participant's S-mSTC ROI, and BOLD activations were calculated as the arithmetic mean of the amplitudes from 4-6 s after stimulus onset (see Figure 3). The activation pattern within S-mSTC showed significant activation only in the 0 ms offset condition (see Figure 3 in blue). Offset levels outside of the range of perceived synchrony were averaged; those greater than 300 ms and defined as asynchronous, showed significant *decreases* in BOLD activation relative to baseline (see Figure 3, in red). Conditions in which the synchrony offset was greater than zero but not outside the range of perceived synchrony were averaged (100-200 ms offsets), but showed no significant difference from baseline (see Figure 3, in grey). Finally, no significant difference was observed between pairwise synchrony effects in S-mSTC across hemispheres ($t = 1.5$).

The second contrast used to identify mSTC was a conjunction of two contrasts using data from the functional localizer runs (in which unisensory blocks of audio-only and visual-only speech were presented). The conjunction of these contrasts, audio presentations > baseline and visual presentations > baseline, defined *bilateral* regions of significant activation in bimodal mSTC (B-mSTC) in each participant (see Figure 6 and Table 1 for average ROI locations). ERAs of BOLD activation from the experimental runs (with the multiple levels of stimulus offset) were extracted from each participant's significant activations within B-mSTC as defined by the functional localizer (see Figure 4), and BOLD activations were calculated as the arithmetic mean of the amplitudes from 4-6 s after stimulus onset. The activation pattern within individually-defined B-mSTC was qualitatively different from the region of mSTC defined by the synchrony-asynchrony contrasts. B-mSTC showed significant activation in all conditions,

with increased activation as the level of asynchrony *increased*. That is, the lowest activation was seen in offsets at or near 0 ms, and the largest activations were seen in the 300-400 ms offset conditions (see Figure 4 in blue and red, respectively). The two patterns of activation showed a significant pairwise difference across B-mSTC and S-mSTC ROIs (pairwise $t = 20.6$; $p < 0.0000001$). B-mSTC did show significantly greater activation with A-V presentations than with V-A presentations as determined by a paired-samples t-test ($t = 4.7$; $p < 0.002$), a finding that mirrors the perceived asynchrony behavioral findings and further suggests that these activations may be related to perceived synchrony. Finally, no significant difference was observed between pairwise synchrony effects in B-mSTC across hemispheres ($t = 0.7$).

This pattern of activation, increasing activation with increasing asynchrony, also co-varied with the total length of stimulus presentation. As stimulus offset increased, the total stimulus presentation time increased to the same extent, creating a possible confound and limiting the interpretation of the activation pattern seen in the individual's ROIs. Our control study accounted for this confound by using all synchronous stimulus presentations with parametrically-varied presentation lengths to match the offsets seen in the initial scan, and performed the same analysis as described above. BOLD activations (and standard errors) observed with increases in presentation time of 0-400 ms, respectively, were 0.08 (0.02), 0.04 (0.02), 0.04 (0.03), 0.05 (0.02), and 0.05 (0.02). No trend was found across stimulus durations, and no significant differences were found between individual duration lengths. Thus, presentation time did not produce any significant changes in BOLD activation within mSTC, ruling out the possibility that the effects reported here were due to slight changes in stimulus presentation length.

In addition to the functional differences observed between individuals' B-mSTC and S-mSTC, the anatomical locations after normalization were found to be significantly different (see Table 1). The mean center of activation in S-mSTC (showing a BOLD activation only when the audiovisual stimulus presentation was synchronized) was located more medially than the mean center of activation in B-mSTC (showing a parametrically increasing BOLD activation with level of asynchrony) in both the left ($p < 0.03$) and right ($p < 0.003$) hemispheres, and was further inferior ($p < 0.003$) in the left hemisphere. Additionally, it should be noted that at this threshold level, the regions of significant activation within these two distinctly defined regions of the mSTC, in addition to showing differing BOLD patterns, were non-overlapping in every participant.

Whole-brain group analysis

To describe a network of brain regions that may be involved in temporal audiovisual processing, experimental runs in which participants were presented with audiovisual spoken words with varying onset asynchronies were analyzed as a group with a balanced contrast comparing synchronous (0 ms offset) and asynchronous (300-400 ms offset) trials (with each asynchronous condition given equal weight). Voxels of activation were deemed significant with a statistical criterion of t-scores 6.00 or greater ($p < 0.004$) with the additional statistical constraint of a cluster-threshold correction of 10 voxels (see Figures 5, 7 (blue), & Table 2), an area of 270 mm^3. The cluster-threshold correction technique used here controls false positives, with a relative sparing of statistical power (Forman et al., 1995; Thirion et al., 2007), and has previously been used to define mSTC (Stevenson et al., 2009). Regions identified in the group data as responding to synchronous multisensory speech signals more than to asynchronous multisensory speech signals included the right mSTC, bilateral LOC, SC, extrastriate visual cortex, and activation within FG extending into the cerebellum (Figure 5). No regions were found in the group analysis that responded more with asynchronous than with synchronous speech signals.

Comparison of individual- and group-defined mSTC

While the whole-brain group analysis (synchronous > asynchronous) produced a significant activation in right S-mSTC, the individual analysis showed *bilateral* S-mSTC activations in all participants. To investigate this discrepancy between group and individual analyses further, localizer runs were also analyzed with a whole-brain, RFX GLM. Consistent with previous studies (Stevenson et al., 2007; Stevenson and James, 2009; Stevenson et al., 2009), mSTC was defined by a conjunction of two contrasts, a contrast of audio presentations > baseline and visual presentations > baseline. A bimodal, unilateral region of the right mSTC (B-mSTC) was identified (see Figure 6a & 7 (orange)), X = 39, Y = −41, Z = 10, t = 26.5, $p < 1.0\times10^{-6}$). This region was located slightly superior to S-mSTC region found with the synchrony-asynchrony contrast. This discrepancy between group-defined and individually-defined ROIs is thus consistent across the synchrony > asynchrony contrast and the conjunction of audio > baseline and visual > baseline contrasts. The discrepancy between group and individual results (see Figure 6) is known to occur (Saxe et al., 2006), and has been reported previously in mSTC and other multisensory regions (Kim and James, In Press; Puce et al., 1998; Stevenson and James, 2009; Stevenson et al., 2009).

Discussion

In everyday processing of sensory information, the human perceptual system shows a remarkable ability to combine multiple sensory signals from a single external event. One of the major sources of information utilized by the perceptual system during audio-visual perceptual fusion in order to identify whether the signals originate from the same source is the degree of temporal synchrony between auditory and visual inputs (Macaluso et al., 2004; Meredith et al., 1987). The importance of temporal synchrony can be clearly seen when temporal processing between modalities is impaired (i.e., as seen with AWF; impairments in the temporal alignment of audiovisual speech produce decreased speech intelligibility (Hamilton et al., 2006)). In the present study, we investigated the brain mechanisms used in such perceptions of asynchronous audiovisual speech in healthy subjects using BOLD fMRI. Two sub-regions of mSTC were identified that showed qualitatively distinct BOLD activation patterns. Synchrony-defined mSTC (S-mSTC) responded in a binary fashion, with significant BOLD activation when the auditory and visual stimuli were synchronous and no activation when stimuli were presented asynchronously (Figure 3). In contrast, bimodal mSTC (B-mSTC) showed significant activation with all presentations, but showed parametrically increasing activation as the level of asynchrony increased (Figure 4).

Multisensory STC regions were defined in each individual according to two distinct functional contrasts. The first contrast, synchrony > asynchrony, identified a bilateral region of mSTC, S-mSTC, that responded more with synchronous than asynchronous speech in every individual (see Table 1 for average ROI locations). Timecourses extracted from individual-defined S-mSTC regions showed a significant activation only with the condition in which the auditory and visual components of the stimulus were synchronous (see Figure 3). That is, even in the cases of 100 ms offsets, there was no increase in BOLD activation over baseline despite the behavioral indications that participants perceived the 100 ms offset conditions as synchronous a high percentage of the time (see Figure 2). Furthermore, such a stimulus-driven activation in S-mSTC, where there is a activation only when there is a synchronous audiovisual presentation, is somewhat contrary to previous findings in mSTC, where there is usually an activation with unisensory-auditory and unisensory-visual stimulus presentations (Beauchamp, 2005a;Beauchamp et al., 2004a;Beauchamp et al., 2004b;Stevenson et al., 2007). Also, the trials in which the stimulus offset was 300 ms or greater elicited a BOLD activation that was less than baseline.

The negative BOLD change observed with asynchronous stimuli in S-mSTC appear contrary to previous findings that have shown that mSTC responds with asynchronous audiovisual presentations (Miller and D'Esposito, 2005) and with both unisensory-auditory or unisensory-visual stimulus presentations (Beauchamp, 2005b; Beauchamp et al., 2004a; Beauchamp et al., 2004b; Stevenson et al., 2007; Stevenson and James, 2009; Stevenson et al., 2009; Werner and Noppeney, In Press). However, when mSTC is defined as a conjunction of regions responding to both unisensory visual-only and auditory-only stimuli (B-mSTC, see Table 1 for average ROI locations), a very different BOLD activation is observed (see Figure 4). Bimodal mSTC, found bilaterally in each individual, produced BOLD activation to any presentation of audiovisual speech, regardless of the level of synchrony. Also, the interaction between BOLD activation and synchrony level in this region was distinct from that in the S-mSTC. While the S-mSTC responded to synchronous speech only (Figure 3), B-mSTC showed parametrically-increasing activation as the level of *asynchrony* increased (Figure 4) -- the greater the temporal offset, the greater the BOLD activation. Furthermore, follow-up scans ruled out the possibility that this effect may have been due to slight increases in total stimulus presentation time occurring on asynchronous trials.

The identification of two anatomically and functionally distinct subregions of mSTC, each showing qualitatively distinct multisensory interactions, suggests that these separate subregions of mSTC are functionally specialized to use different neural mechanisms to integrate auditory and visual information in speech. Our results suggest that S-mSTC is sensitive to external-stimulus synchrony, responding only when the two stimulus components are temporally aligned. This may reflect a synchrony-detection process facilitating perceptual fusion only when the auditory and visual signals are temporally synchronous. Previous effective-connectivity research using non-speech stimuli has suggested that temporal coincidence of an auditory and visual input resulted in mSTC activity showing a greater influence on primary auditory and visual cortices (Noesselt et al., 2007). Our finding of a possible synchrony-detecting subregion of mSTC may provide a processing mechanism for these effects. Another possibility is that S-mSTC acts as a first pass filter: if the multisensory stimulus pair is synchronous then it is perceptually fused, and if not, the signal is further processed, perhaps in B-mSTC.

The B-mSTC subregion showed greater activation when larger stimulus offsets were present in the multisensory stimulus. A number of possible neuronal processing mechanisms may account for such a phenomenon. First, facilitation in the BOLD activation may be due to an increase in the level or duration of neural activity reflecting an increase in processing demands during integration of asynchronous speech (Formisano et al., 2002; Georgopoulos et al., 1989; Miller and D'Esposito, 2005; Richter et al., 2000). A second possibility is that B-mSTC responds to a greater extent when an individual perceives multiple sensory inputs. That is, when the audio and visual inputs are fused, this region processes one percept, whereas when the audio and visual inputs are not perceptually fused, the region processes two separate percepts, requiring an increase in neural processing. As the stimulus presentation in this experiment became more asynchronous, the percentage of trials in which the participants failed to fuse the auditory and visual components increased, which may account for the parametric increase in BOLD activation with stimulus asynchrony. The current experimental design, however, is not able to distinguish between these competing hypotheses. These aforementioned explanations are not mutually exclusive and warrant further investigation.

Another avenue of research that remains unresolved here is the manner in which these subregions of mSTC interact. Previous studies using similar functional definitions have suggested that mSTC may be functionally heterogeneous, in particular, that there may be a patchy organization of subregions with small unisensory subregions that feed forward into integrative subregions (Beauchamp et al., 2004a). The functional connectivity between B-

mSTC and S-mSTC is unclear, and future studies would be needed before any claims about connectivity could be made, but the present findings do suggest that processes related to temporal synchrony perception and perceptual fusion impairments may be distinct and utilize separate brain regions.

The ROIs described above are only two nodes in a larger network that is involved in the processing of audiovisual speech. Thus, in addition to the individual subject ROI analysis, we conducted a whole-brain group-averaged SPM analysis contrasting BOLD activation with synchronous (0 ms offset) and asynchronous (300-400 ms offset) speech to describe the network of brain regions sensitive to temporal congruency of auditory and visual stimuli. The synchrony-sensitive network included bilateral SC, bilateral LO, bilateral extrastriate visual cortex, right-lateralized mSTC, and a right-lateralized activation in both the fusiform gyrus and cerebellum (Figure 5 and Table 2). The network included regions that have been previously shown to be sensitive to temporal offsets as well as novel regions.

Two of these regions, SC and mSTC, are commonly studied in association with audiovisual integration. Single-unit recordings from SC in a number of mammals have shown that multisensory cells respond to a greater extent when the multisensory stimuli are temporally synchronous (King and Palmer, 1985; Meredith et al., 1987; Stein et al., 1993). Functional MRI studies of humans have also shown that human SC exhibits an enhanced activation when audiovisual stimuli are presented relative to unisensory audio-only and visual-only presentations (Calvert et al., 2000). What has been less clear, however, is the effect that temporal synchrony has on multisensory interactions within human SC. Calvert and colleagues (2000) showed that speech stimuli that were both temporally and semantically congruent produced a superadditive activation in SC, whereas temporally and semantically incongruent speech produced a sub-additive SC activation. Macaluso and colleagues (2004) compared activation with temporal offsets of a multisensory stimulus set, but reported no effect in the SC. Miller and D'Esposito (2005) also used asynchronous audiovisual speech, but reported an enhanced activation in the SC when stimuli were asynchronous, the exact opposite finding commonly reported in single-unit studies and other imaging studies (Calvert et al., 2000)[1]. Our findings here report an interaction with temporal congruency similar to the single-unit recordings, that is, we found a greater activation in SC with synchronous relative to asynchronous presentations of multisensory speech signals.

Like the findings concerning synchrony effects in SC, BOLD activation in mSTC has been inconsistent, with some studies reporting no synchrony effect (Miller and D'Esposito, 2005) and some reporting only unilateral effects in the left hemisphere (Macaluso et al., 2004). In the group SPM reported here, we also identified a unilateral region of mSTC that showed a synchrony effect (S-mSTC), but this region was *right*-lateralized.

While we have focused on differences between S-mSTC and B-mSTC up to this point, there is also a noteworthy similarity between the subregions, both S-mSTC and B-mSTC were only identified *unilaterally* in the RFX group analysis, yet were both identified *bilaterally* when ROIs were defined on an individual-by-individual basis. This observation is not unique. Previous reports have also shown significant differences between group-defined mSTC and individually-defined mSTC (Stevenson et al., 2007; Stevenson and James, 2009; Stevenson et al., 2009; Wheaton et al., 2004), an effect that has largely been explained as resulting from substantial variability in the anatomical location of individual's functionally-defined mSTC (Stevenson and James, 2009). Our results provide additional support for this hypothesis, as can be seen most clearly in the comparison of group-defined and individual-defined B-mSTC (see

[1]It should also be noted that imaging of the SC is often difficult due to susceptibility artifacts because of its anatomical location, which may contribute to some discrepancies between fMRI studies.

Figure 6). A simple group RFX GLM analysis showed a strong right-lateralized activation with no corresponding left-hemisphere activation (Figure 6a). The individual analysis, however, revealed that *bilateral* activations in every individual. The anatomical locations of the functionally-defined individual ROIs in the right hemisphere are much more spatially homogenous in than the left hemisphere. This difference in function-to-structure homogeneity across individuals observed here, taken together with the RFX GLM that is commonly used with group analyses, seems a likely source of these differences found between group and individual results. As a result, it is advisable to incorporate both of these forms of analyses to utilize the benefits of each.

One advantage of parametrically varied additive factors is that this method provides a means to not only identify multisensory brain regions, but also to make inferences about the type of stimulus information modulating the integrative process occurring within a given region (James et al., 2009; Stevenson et al., 2009; Werner and Noppeney, In Press). Such an additive-factors design in BOLD fMRI compares the difference in BOLD signal with multisensory stimuli across several levels of a given added factor (ΔAV) with the summed differences in BOLD signal with the unisensory components across the same levels of the added factor ($\Delta A + \Delta V$). An inequality between the multisensory difference and the summed unisensory difference ($\Delta AV \neq \Delta A + \Delta V$) indicates an interaction, suggesting that the stimulus information associated with the added factor is being processed in that region (Stevenson et al., 2009). This may be of particular importance in multisensory audiovisual speech integration because there are a number of reliable effects that are produced by modulating specific stimulus factors. For example, temporal asynchrony and spatial congruency can be modulated to identify networks associated with the ventriloquist effect (Bertelson et al., 2000a; Bertelson and Radeau, 1981; Bertelson et al., 2000b; Stekelenburg et al., 2004; Vroomen et al., 2001) or stimulus quality may be modulated to investigate multisensory enhancement at low signal-to-noise ratios (James et al., 2009; Kim and James, In Press; Stevenson and James, 2009; Sumby and Pollack, 1954; Summerfield, 1987; Werner and Noppeney, In Press).

In the present study, we used temporal synchrony as an added factor. The modulation of synchrony in the unisensory conditions produced a BOLD difference of zero ($\Delta A = 0$, $\Delta V = 0$), resulting in any non-zero value of ΔAV (as seen in the synchrony > asynchrony contrast) indicating an interaction. This analysis identifying a synchrony-sensitive network including SC, mSTC, LO, FG, and extrastriate visual cortex, can then be compared to other networks showing interactions across changes in other added factors. For example, integrative networks influenced by parametric manipulations of stimulus quality have been identified in the same manner that we have here identified a synchrony-sensitive network (Stevenson et al., 2009). This synchrony-sensitive network includes a number of regions that overlap with the previously-described network that was found to be sensitive to stimulus quality, including right mSTC. This overlap suggests that mSTC differentially integrates audiovisual speech signals according to both temporal information and stimulus quality, and thus may be involved in behavioral effects such as the ventriloquist effect (Bertelson and Radeau, 1981) and inverse effectiveness (Stevenson and James, 2009; Stevenson et al., 2009; Werner and Noppeney, In Press).

While network overlaps are informative, non-overlapping regions in the synchrony-sensitive and stimulus-quality-sensitive networks can also be informative. Integration in LO and right FG, for example, is modulated by temporal synchrony but not stimulus quality. Likewise, medial frontal gyrus, caudate nucleus, and posterior cingulate gyrus were modulated by stimulus quality but not temporal synchrony (Stevenson et al., 2009). These differences in sensitivity to individual stimulus factors reflect similar behavioral multisensory effects in speech perception. Perceptual fusion of auditory and visual information streams is dependent upon temporal synchrony to the extent that temporal synchrony can often override a spatial

discrepancy, as seen in the Ventriloquist effect (Bertelson et al., 2000a; Bertelson and Radeau, 1981; Bertelson et al., 2000b; Stekelenburg et al., 2004; Vroomen et al., 2001), despite the findings that perceptual fusion of speech is highly robust against small changes in temporal synchrony (Conrey and Pisoni, 2004; Dixon and Spitz, 1980; Grant and Greenberg, 2001; Grant et al., 2004; McGrath and Summerfield, 1985; Miller and D'Esposito, 2005; van Wassenhove et al., 2007). Also, the level of stimulus quality in speech perception has been shown to modulate the gain seen when faces and voices are combined (Ross et al., 2007; Stevenson and James, 2009; Sumby and Pollack, 1954). The presence of distinct but overlapping neural networks that are reflective of these behavioral effects, being sensitive to either temporal offset or stimulus quality, suggests that these two networks are differentially involved in multisensory integration processes.

It should be noted that this comparison of neural networks is based on analysis across subjects and across studies, and as such, should be taken as preliminary evidence. Future comparisons of such networks may prove fruitful, particularly studies directly comparing networks within the same experiment and utilizing concurrent behavioral measures of the multisensory effect of interest, or multisensory impairments in clinical cases similar to AWF.

Conclusions

This study identified functionally and anatomically distinct subregions within mSTC that exhibit qualitatively different BOLD activation patterns in response to changes in the synchrony level of audiovisual speech signals, suggesting that they may sub-serve different integrative processes. S-mSTC responded only to synchronous audiovisual stimulus presentations, with no activation observed at asynchrony levels greater than or equal to 100 ms, while B-mSTC showed activation with any multisensory audiovisual speech stimulus, and increased activation as stimulus asynchrony increases. Additionally, these findings suggest that other subregions of mSTC may be processing other stimulus properties that are important for multisensory integration, such as spatial and semantic congruency. The presence of multiple sites of multisensory integration that differentially process audiovisual speech also suggests that effects related to perceptual fusion and perceptual synchrony may be independent in some ways, although the two are strongly related.

Acknowledgments

This research was supported in part by the Indiana METACyt Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc., by a grant to T. W. James from Indiana University's Faculty Research Support Program administered by the office of the vice provost for research, NIH NIDCD Training grant T32 DC000012 Training in Speech, Hearing, and Sensory Communication, NIH NIDCD Research Grant R01 DC-00111, and the Indiana University GPSO Research Grant. Thanks to Laurel Stevenson, Youngsuk Altieri, June Young Lee, Beth Greene, and Karin Harman James for their support, to Luis Hernandez for the stimuli, and the Indiana University Neuroimaging Group for their insights on this work.

References

Beauchamp MS. See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. Curr Opin Neurobiol 2005a;15:145–153. [PubMed: 15831395]

Beauchamp MS. Statistical criteria in FMRI studies of multisensory integration. Neuroinformatics 2005b; 3:93–113. [PubMed: 15988040]

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci 2004a;7:1190–1192. [PubMed: 15475952]

Beauchamp MS, Lee KE, Argall BD, Martin A. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 2004b;41:809–823. [PubMed: 15003179]

Bertelson P, Pavani F, Ladavas E, Vroomen J, de Gelder B. Ventriloquism in patients with unilateral visual neglect. Neuropsychologia 2000a;38:1634–1642. [PubMed: 11074086]

Bertelson P, Radeau M. Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. Percept Psychophys 1981;29:578–584. [PubMed: 7279586]

Bertelson P, Vroomen J, de Gelder B, Driver J. The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept Psychophys 2000b;62:321–332. [PubMed: 10723211]

Brainard DH. The Psychophysics Toolbox. Spat Vis 1997;10:433–436. [PubMed: 9176952]

Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol 2000;10:649–657. [PubMed: 10837246]

Conrey B, Pisoni DB. Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. J Acoust Soc Am 2006;119:4065–4073. [PubMed: 16838548]

Conrey, BL.; Pisoni, DB. Detection of Auditory-Visual Asynchrony in Speech and Nonspeech Signals.. In: Pisoni, DB., editor. Research on Spoken Language Processing. Indiana University; Bloomington: 2004. p. 71-94.

Dixon NF, Spitz L. The detection of auditory visual desynchrony. Perception 1980;9:719–721. [PubMed: 7220244]

Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn Reson Med 1995;33:636–647. [PubMed: 7596267]

Formisano E, Linden DE, Di Salle F, Trojano L, Esposito F, Sack AT, Grossi D, Zanella FE, Goebel R. Tracking the mind's image in the brain I: time-resolved fMRI during visuospatial mental imagery. Neuron 2002;35:185–194. [PubMed: 12123618]

Gaver WW. What in the world do we hear?: An ecological approach to auditory event perception. Ecological Psychology 1993;5:1–29.

Georgopoulos AP, Lurito JT, Petrides M, Schwartz AB, Massey JT. Mental rotation of the neuronal population vector. Science 1989;243:234–236. [PubMed: 2911737]

Glover GH. Deconvolution of impulse response in event-related BOLD fMRI. Neuroimage 1999;9:416–429. [PubMed: 10191170]

Grant, KW.; Greenberg, S. Speech intelligibility derived from asynchronous processing of auditory-visual information.. International Conference of Auditory-Visual Speech Processing; Santa Cruz, CA. 2001. p. 132-137.

Grant KW, Van Wassenhove V, Poeppel D. Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. Speech Communication 2004;44:43–53.

Grant KW, Walden BE, Seitz PF. Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. J Acoust Soc Am 1998;103:2677–2690. [PubMed: 9604361]

Green KP, Kuhl PK. The role of visual information in the processing of place and manner features in speech perception. Percept Psychophys 1989;45:34–42. [PubMed: 2913568]

Hamilton RH, Shenton JT, Coslett HB. An acquired deficit of audiovisual speech processing. Brain Lang 2006;98:66–73. [PubMed: 16600357]

James, TW.; Stevenson, RA.; Kim, S. The International Society for Psychophysics. Dublin, Ireland: 2009. Assessing multisensory integration with additive factors and functionsl MRI..

Kim S, James TW. Enhanced Effectiveness in visuo-haptic object-selective brain regions with increasing stimulus saliency. Hum Brain Mapp. In Press.

King AJ, Palmer AR. Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. Exp Brain Res 1985;60:492–500. [PubMed: 4076371]

Lachs, L.; Hernandez, LR. Update: The Hoosier Audiovisual Multitalker Database.. In: Pisoni, DB., editor. Research on spoken language processing. Speech Research Laboratory, Indiana University; Bloomington, IN: 1998. p. 377-388.

Laurienti PJ, Kraft RA, Maldjian JA, Burdette JH, Wallace MT. Semantic congruence is a critical factor in multisensory behavioral performance. Exp Brain Res 2004;158:405–414. [PubMed: 15221173]

Luce PA, Pisoni DB. Recognizing spoken words: the neighborhood activation model. Ear Hear 1998;19:1–36. [PubMed: 9504270]

Macaluso E, George N, Dolan R, Spence C, Driver J. Spatial and temporal factors during processing of audiovisual speech: a PET study. Neuroimage 2004;21:725–732. [PubMed: 14980575]

McGrath M, Summerfield Q. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. J Acoust Soc Am 1985;77:678–685. [PubMed: 3973239]

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature 1976;264:746–748. [PubMed: 1012311]

Meredith MA, Nemitz JW, Stein BE. Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. J Neurosci 1987;7:3215–3229. [PubMed: 3668625]

Miller LM, D'Esposito M. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. J Neurosci 2005;25:5884–5893. [PubMed: 15976077]

Munhall KG, Gribble P, Sacco L, Ward M. Temporal constraints on the McGurk effect. Percept Psychophys 1996;58:351–362. [PubMed: 8935896]

Noesselt T, Rieger JW, Schoenfeld MA, Kanowski M, Hinrichs H, Heinze HJ, Driver J. Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. J Neurosci 2007;27:11431–11441. [PubMed: 17942738]

Pelli DG. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis 1997;10:437–442. [PubMed: 9176953]

Puce A, Allison T, Bentin S, Gore JC, McCarthy G. Temporal cortex activation in humans viewing eye and mouth movements. J Neurosci 1998;18:2188–2199. [PubMed: 9482803]

Richter W, Somorjai R, Summers R, Jarmasz M, Menon RS, Gati JS, Georgopoulos AP, Tegeler C, Ugurbil K, Kim SG. Motor area activity during mental rotation studied by time-resolved single-trial fMRI. J Cogn Neurosci 2000;12:310–320. [PubMed: 10771414]

Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex 2007;17:1147–1153. [PubMed: 16785256]

Saxe R, Brett M, Kanwisher N. Divide and conquer: a defense of functional localizers. Neuroimage 2006;30:1088–1096. discussion 1097-1089. [PubMed: 16635578]

Scott SK, Blank CC, Rosen S, Wise RJ. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 2000;123(Pt 12):2400–2406. [PubMed: 11099443]

Sheffert, SM.; Lachs, L.; Hernandez, LR. The Hooiser Audiovisual Multitalker Database.. In: Pisoni, DB., editor. Research on spoken language processing. Speach Research Laboratory, Indiana University; Bloomington, IN: 1996. p. 578-583.

Soto-Faraco S, Alsius A. Deconstructing the McGurk-MacDonald illusion. J Exp Psychol Hum Percept Perform 2009;35:580–587. [PubMed: 19331510]

Stein BE, Meredith MA, Wallace MT. The visually responsive neuron and beyond: multisensory integration in cat and monkey. Prog Brain Res 1993;95:79–90. [PubMed: 8493355]

Steinmetz R. Human perception of jitter and media synchronization. IEEE Journal on Selected Areas in Communication 1996;14:61–72.

Stekelenburg JJ, Vroomen J, de Gelder B. Illusory sound shifts induced by the ventriloquist illusion evoke the mismatch negativity. Neurosci Lett 2004;357:163–166. [PubMed: 15003275]

Stevenson RA, Geoghegan ML, James TW. Superadditive BOLD activation in superior temporal sulcus with threshold non-speech objects. Experimental Brain Research 2007;179:85–95.

Stevenson RA, James TW. Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. Neuroimage 2009;44:1210–1223. [PubMed: 18973818]

Stevenson RA, Kim S, James TW. 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 2009;198:183–194. [PubMed: 19352638]

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 1954;26:212–215.

Summerfield, Q. Some Preliminaries to a Comprehensive Account of Audio-visual Speech Perception.. In: Dodd, B.; Campbell, BA., editors. Hearing by Eye: The Psychology of Lip Reading. Lawrence Erlbaum Associates Ltd.; Publishers, London, UK: 1987. p. 3-52.

Talaraich, J.; Tournoux, P. Co-planar stereotaxic atlas of the human brain. Theime Medical Publishers; New York, New York: 1988.

Thirion B, Pinel P, Meriaux S, Roche A, Dehaene S, Poline JB. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. Neuroimage 2007;35:105–120. [PubMed: 17239619]

van Atteveldt NM, Formisano E, Blomert L, Goebel R. The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. Cereb Cortex 2007;17:962–974. [PubMed: 16751298]

van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci U S A 2005;102:1181–1186. [PubMed: 15647358]

van Wassenhove V, Grant KW, Poeppel D. Temporal window of integration in auditory-visual speech perception. Neuropsychologia 2007;45:598–607. [PubMed: 16530232]

Vroomen J, Bertelson P, de Gelder B. The ventriloquist effect does not depend on the direction of automatic visual attention. Percept Psychophys 2001;63:651–659. [PubMed: 11436735]

Werner S, Noppeney U. Superadditive responses in the superior temporal sulcus predict audiovisual benefits in object categorization. Cerebral Cortex. In Press.

Wheaton KJ, Thompson JC, Syngeniotis A, Abbott DF, Puce A. Viewing the motion of human body parts activates different regions of premotor, temporal, and parietal cortex. Neuroimage 2004;22:277–288. [PubMed: 15110018]

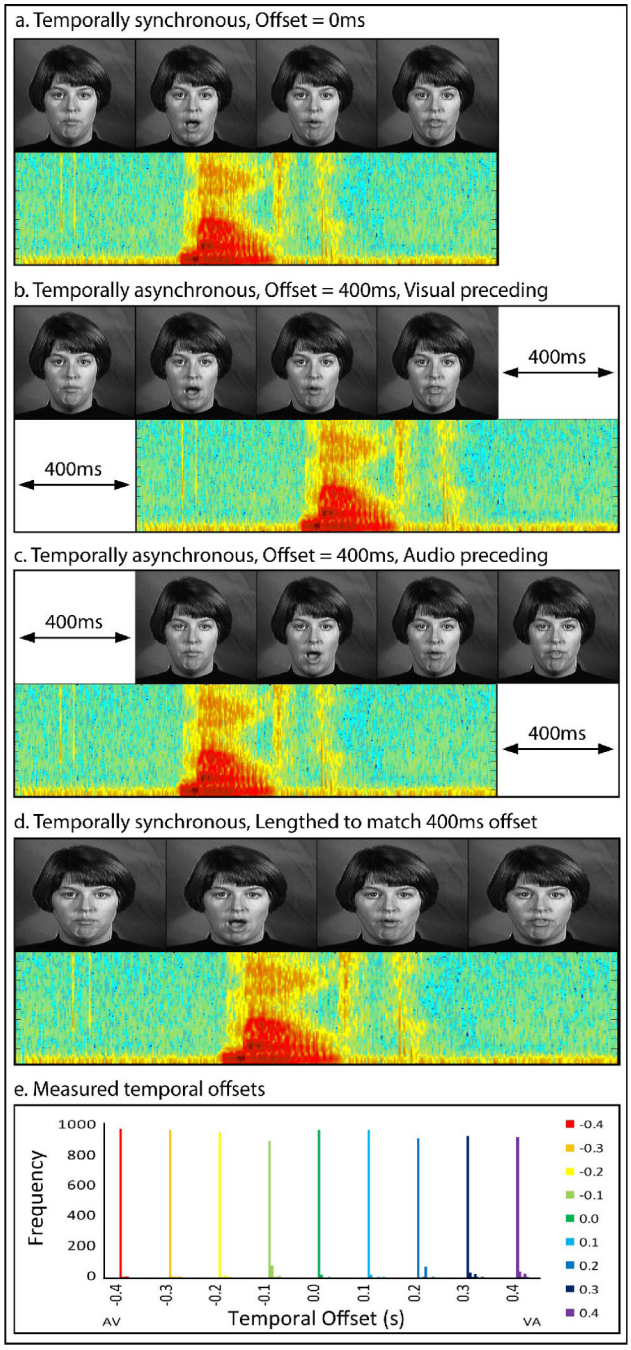

Figure 1. Example Stimuli

Stimulus presentations included synchronous auditory and visual components (a) and asynchronous presentations with offsets ranging from 100 to 400 ms with both visual (b) and auditory (c) components presented first. In an additional paradigm run to account for stimulus presentation time, visual frame rates and auditory bits per second were increased to match the asynchronous stimulus presentation times (d). In order to ensure timing precision, 1000 timing offsets were measured for each stimulus asynchrony level, (with different colors representing each offset condition), and frequencies out of 1000 were calculated in 5 ms bins (e).

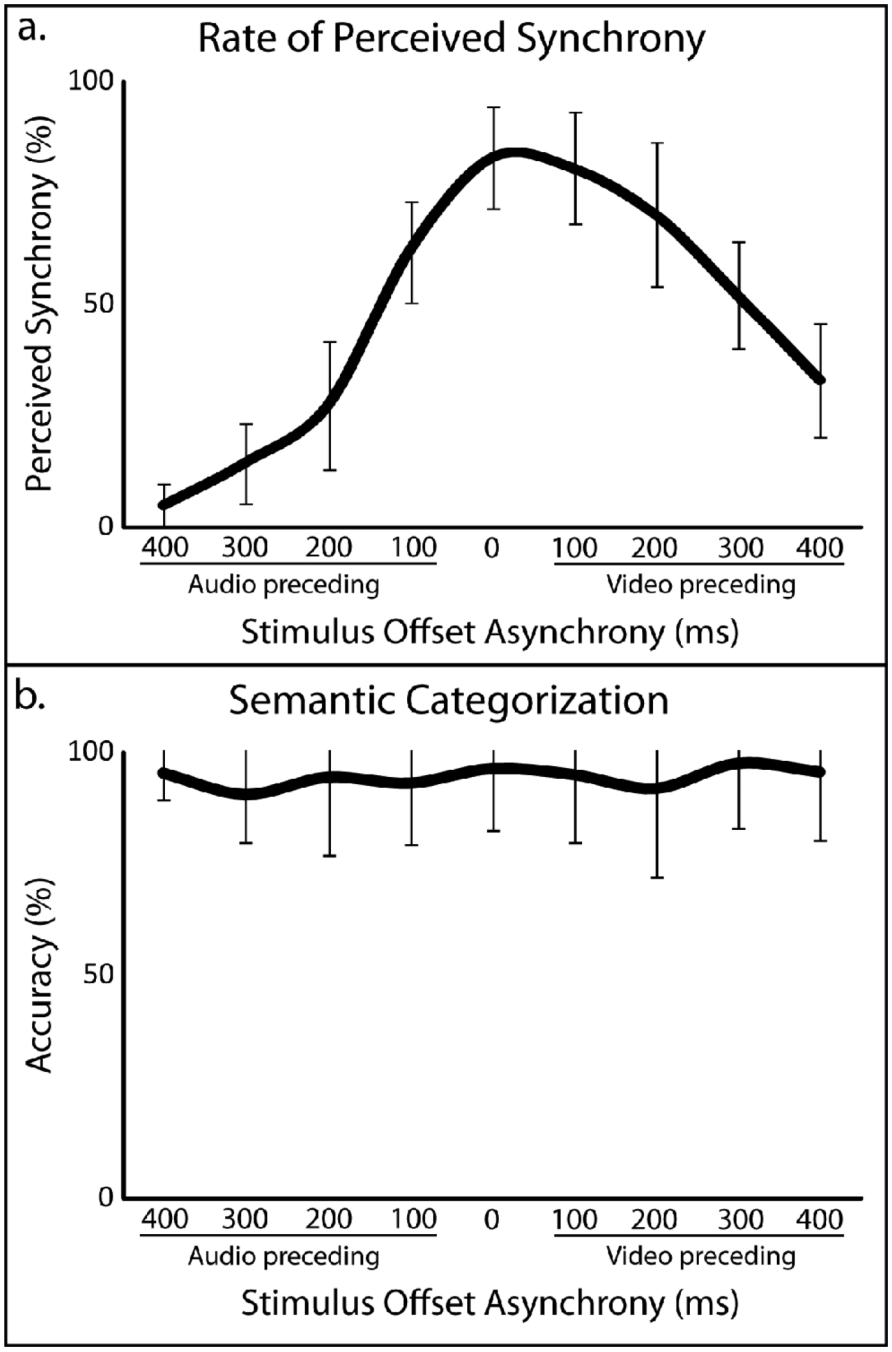

Figure 2. Behavioral results

Perceived synchrony rates were collected during a pre-scan behavioral session (a). Accuracy rates of a two-alternative, forced-choice semantic categorization task during fMRI scanning sessions (b). Error bars in both panels reflect between-subject standard deviations.

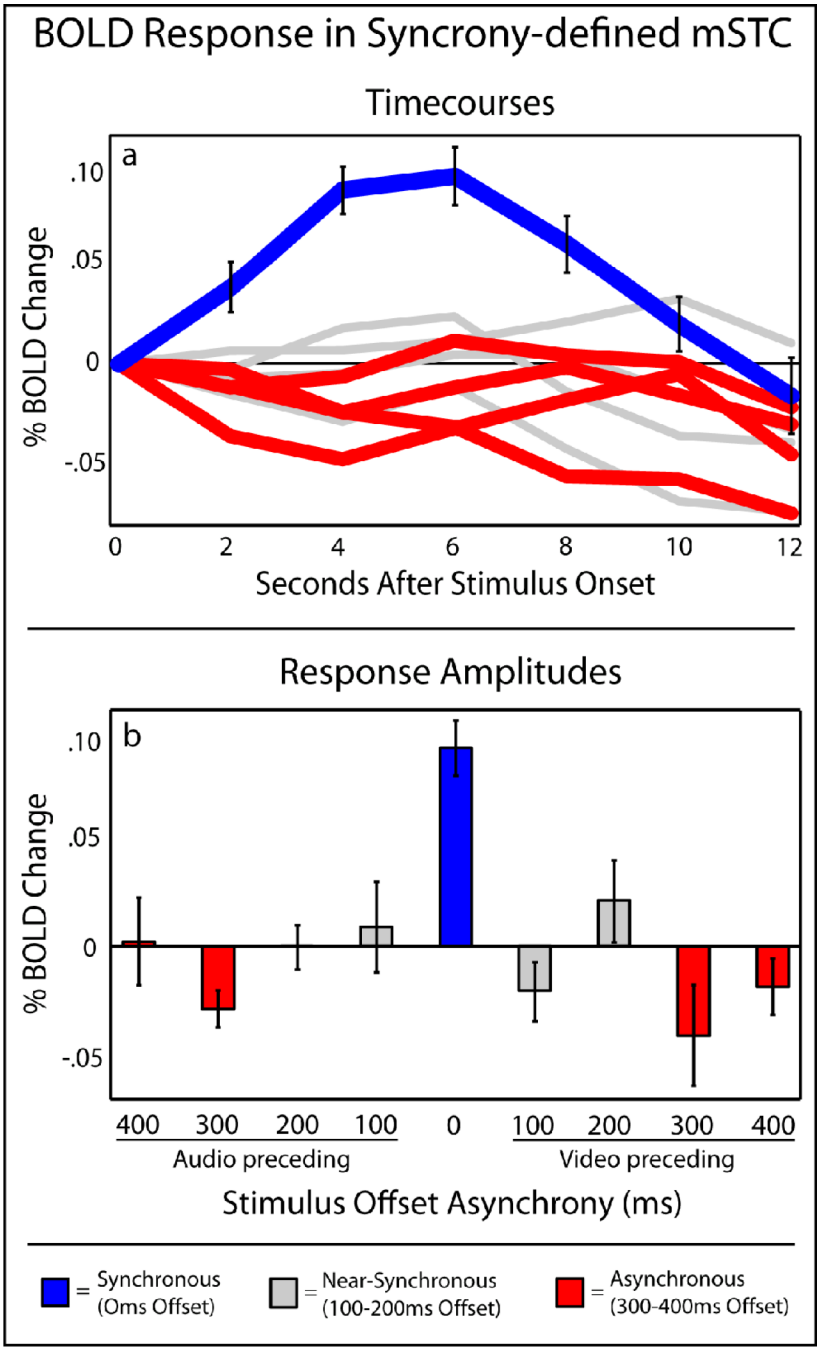

Figure 3. BOLD responses in S-mSTC

Averaged timecourses (a) and BOLD response amplitudes (b) across levels of asynchrony extracted from individual subject's synchrony-defined mSTC ROIs, as defined by a synchronous (blue) > asynchronous (red) contrast.

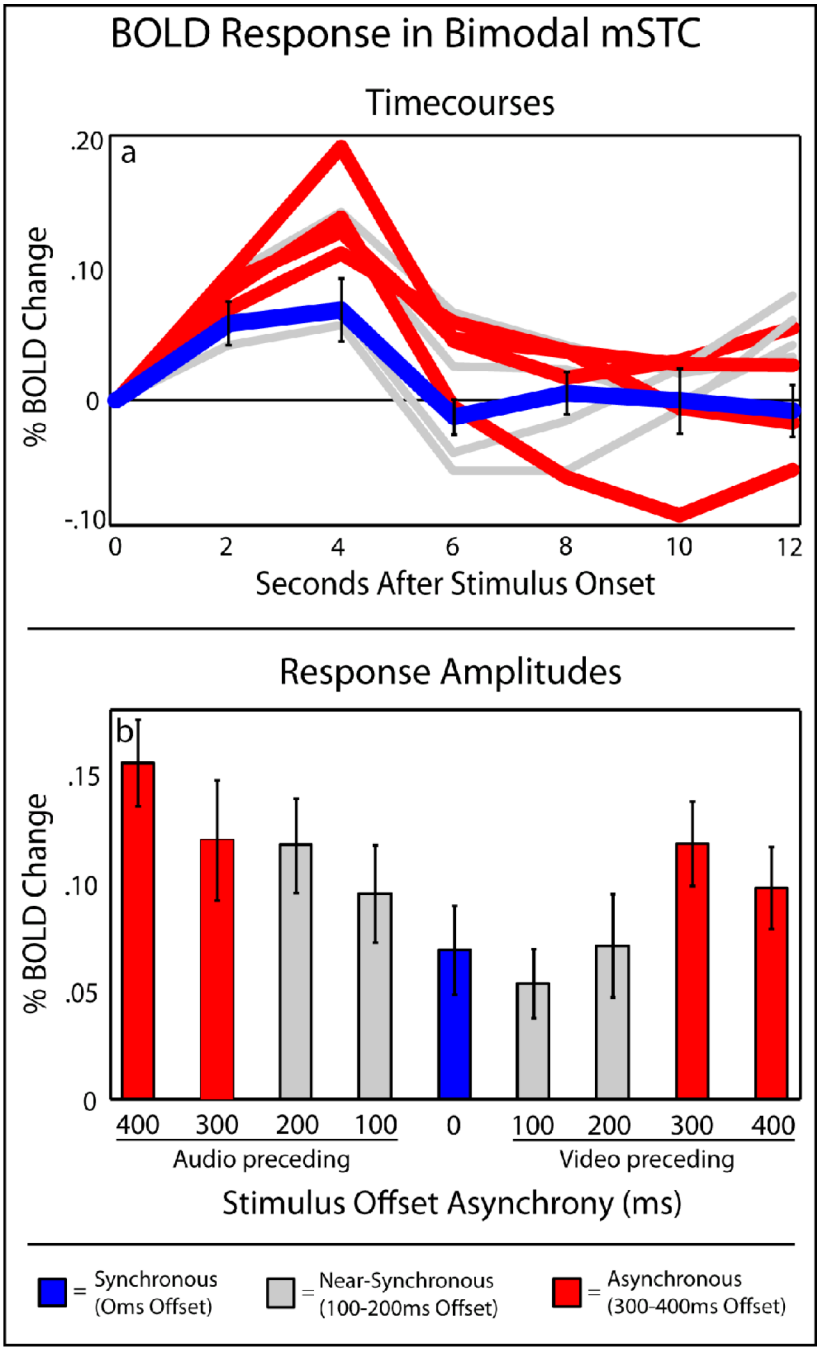

Figure 4. BOLD responses in B-mSTC

Averaged timecourses (a) and BOLD response amplitudes (b) across levels of asynchrony extracted from individual subject's bimodal mSTC ROIs, as defined by a conjunction of activations with unisensory-auditory and unisensory-visual stimulus presentations. Synchronous and asynchronous trials used to define the other sub-region of mSTC are labeled in blue and red, respectively, for comparison.

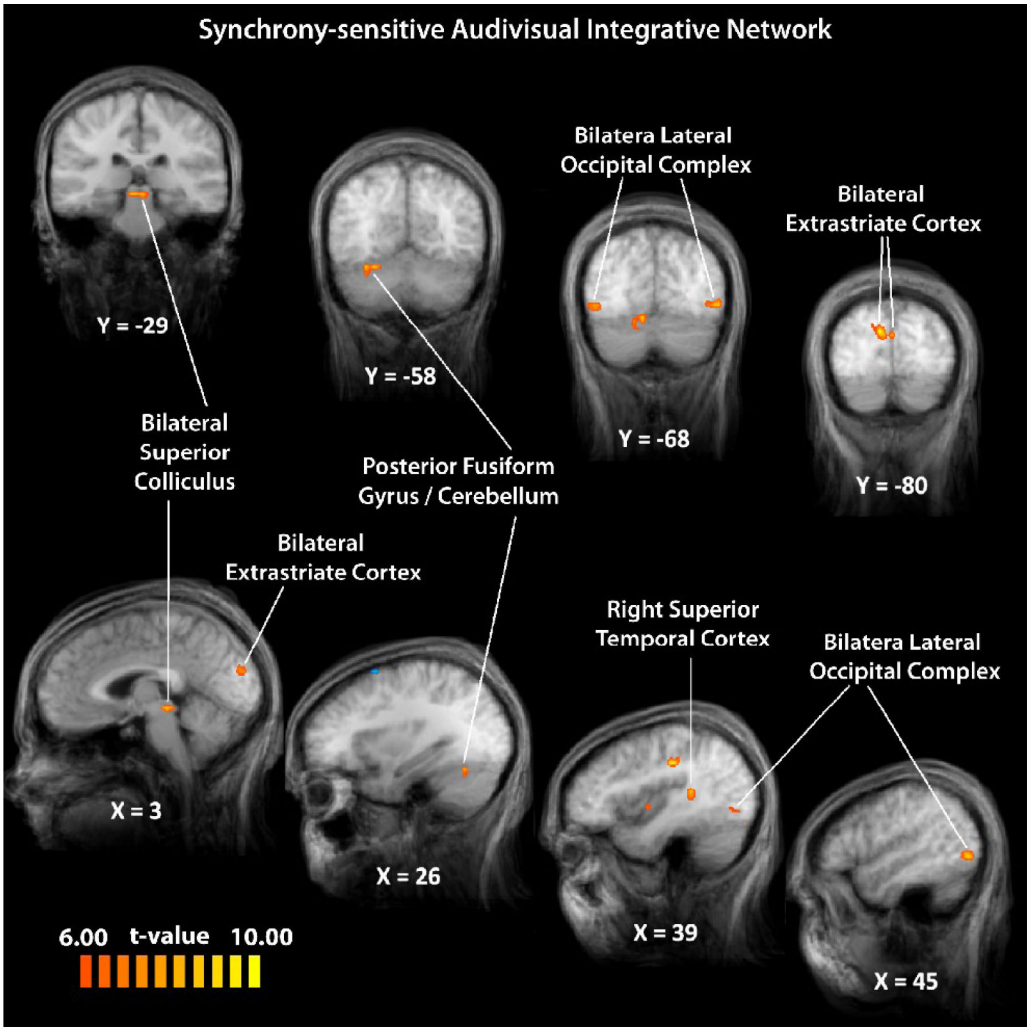

Figure 5. A synchrony-sensitive integrative network

Using a synchronous > asynchronous contrast, a network of regions sensitive to modulations of temporal synchrony was identified.

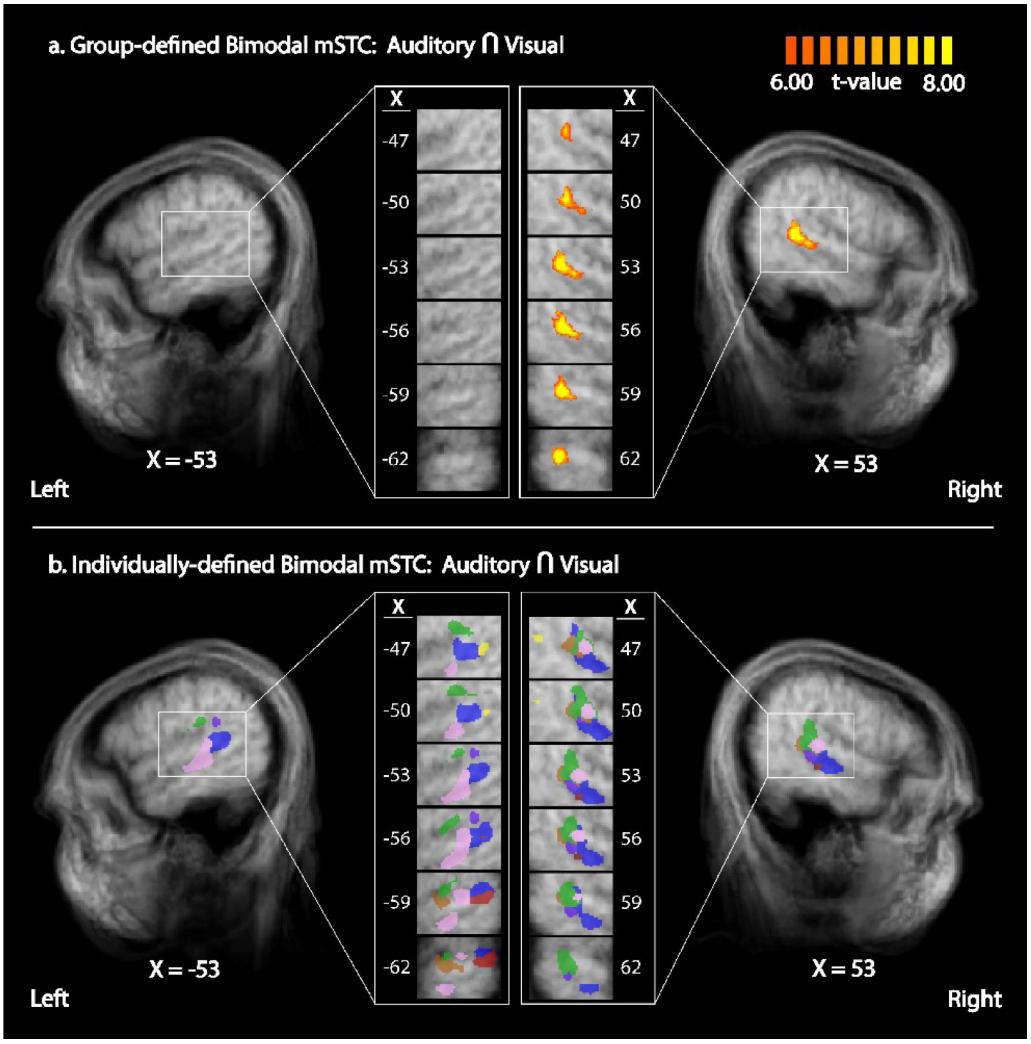

Figure 6. Bimodal mSTC defined by group and individual GLMs

Bimodal mSTC was defined as the conjunction of regions that exhibited significant activation with both unisensory-auditory and unisensory-visual stimulus. A group contrast a revealed *unilateral* activation in right B-mSTC (a). Individual analysis using the same contrast revealed *bilateral* B-mSTC activation (b). Right individually-defined ROIs showed greater anatomical homogeneity (with each color representing a different individual) compared to left ROIs, a result that may explain the lack of significant group activation in left mSTC.

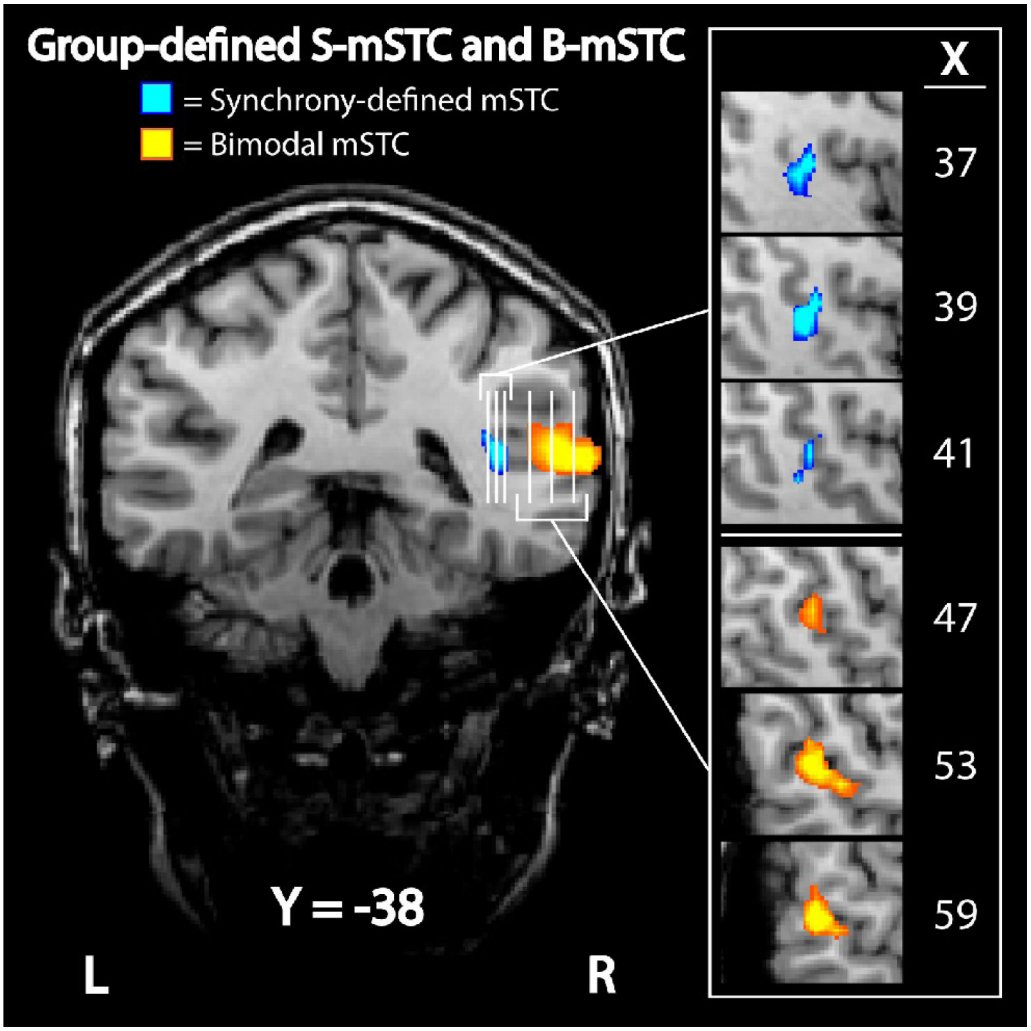

Figure 7. Group-defined subregions of mSTC

Two distinct regions of mSTC were defined with group data. The first subregion, S-mSTC, was defined with a synchronous > asynchronous contrast (blue), while the second subregion, B-mSTC, was defined using a conjunction of two contrasts, audio-only > baseline and visual-only > baseline (orange). Each vertical white line on the coronal image represents a slice that can be seen in the sagittal orientation to the right.

Table 1

mSTC Regions defined by individual's ROIs

ROI definition	Hemisphere	X	Y	Z	Voxels	t	p
Synchrony (S-mSTC)	Right	36(1.4)*†	−37(2.1)	8(3.0)	129	3.0	< 0.003
Synchrony (S-mSTC)	Left	−45(2.2)*†	−43(4.1)*	1(3.0)*	235	3.0	< 0.003
A ∩ V (B−mSTC)	Right	53(1.4)*	−36(3.5)	8(2.1)	1079	6.5	< $6.00e^{-10}$
A ∩ V (B−mSTC)	Left	−54(2.4)*	−35(3.4)*	16(2.8)*	1065	6.5	< $6.00e^{-10}$

Talaraich coordinates reported as mean(standard errors).

*= significant difference between synchrony and A ∩ V defined ROIS

†= significant difference between left and right hemispheres

Table 2

Regions defined in group RFX SPM: Synchronous - Asynchronous

Region	Hemisphere	X	Y	Z	Voxels	t*	p*
Superior Temporal Cortex	Right	39	−37	13	170	9.5	$1.50e^{-5}$
Posterior Fusiform Gyrus / Cerebellum	Right	26	−58	−15	192	12.6	$2.32e^{-6}$
Superior Colliculus	Right	6	−29	−4	411	6.7	$1.39e^{-4}$
Superior Colliculus	Left	−2	−28	−5	246	6.7	$1.39e^{-4}$
Lateral Occipital Complex	Right	46	−68	0	447	8.0	$4.55e^{-5}$
Lateral Occipital Complex	Left	−44	−67	3	556	11.8	$3.56e^{-6}$
Extrastriate Visual Cortex	Right	10	−78	24	663	12.5	$2.48e^{-6}$
Extrastriate Visual Cortex	Left	−1	−82	23	344	11.5	$4.34e^{-6}$

*t-values and associated p-values reported are the uncorrected maximum values within a region of interest defined using a cluster threshold to control for false positives