Perception drives production across sensory modalities: A network for sensorimotor integration of visual speech

Jonathan H. Venezia[1], **Paul Fillmore**[2], **William Matchin**[3], **A. Lisette Isenberg**[1], **Gregory Hickok**[1], and **Julius Fridriksson**[4]

[1]Department of Cognitive Sciences; University of California, Irvine; Irvine, CA, 92697; United States

[2]Department of Communication Sciences and Disorders; Baylor University; Waco, TX, 76798; United States

[3]Department of Linguistics; University of Maryland; College Park, MD, 20742; United States

[4]Department of Communication Sciences and Disorders; University of South Carolina; Columbia, SC, 29208; United States

Abstract

Sensory information is critical for movement control, both for defining the targets of actions and providing feedback during planning or ongoing movements. This holds for speech motor control as well, where both auditory and somatosensory information have been shown to play a key role. Recent clinical research demonstrates that individuals with severe speech production deficits can show a dramatic improvement in fluency during online mimicking of an audiovisual speech signal suggesting the existence of a *visuomotor* pathway for speech motor control. Here we used fMRI in healthy individuals to identify this new visuomotor circuit for speech production. Participants were asked to perceive and covertly rehearse nonsense syllable sequences presented auditorily, visually, or audiovisually. The motor act of rehearsal, which is *prima facie* the same whether or not it is cued with a visible talker, produced different patterns of sensorimotor activation when cued by visual or audiovisual speech (relative to auditory speech). In particular, a network of brain regions including the left posterior middle temporal gyrus and several frontoparietal sensorimotor areas activated more strongly during rehearsal cued by a visible talker versus rehearsal cued by auditory speech alone. Some of these brain regions responded exclusively to rehearsal cued by visual or audiovisual speech. This result has significant implications for models of speech motor control, for the treatment of speech output disorders, and for models of the role of speech gesture imitation in development.

Corresponding Author: Jonathan Venezia, 3151 Social Science Plaza, University of California, Irvine, Irvine, CA 92697, Phone: (949) 824-1409, Fax: (949) 824-2307, jvenezia@uci.edu.

1. Introduction

Visual speech – the time-varying and pictorial cues associated with watching a talker's head, face and mouth during articulation – has a well-known and dramatic effect on speech perception. Adding visual speech to an auditory speech signal improves intelligibility for both clear (Arnold and Hill, 2001) and distorted (Erber, 1969; MacLeod and Summerfield, 1987; McCormick, 1979; Neely, 1956; Ross et al., 2007; Sumby and Pollack, 1954) speech, improves acquisition of non-native speech sound categories (Hardison, 2003), and can alter the percept of an acoustic speech stimulus (McGurk and MacDonald, 1976). It is no surprise, then, that the vast majority of research and theory involving visual speech concerns its influence on perception. Conversely, research on auditory speech has focused not only on its fundamental contribution to perception, but also on the crucial role of auditory speech systems in supporting speech production. Specifically, evidence suggests that auditory speech representations serve as the sensory targets for speech production. Briefly, auditory speech targets are predominantly-syllabic representations in the posterior superior temporal lobe that comprise the higher-level (i.e., linguistic) sensory goals for production (Hickok, 2012). Articulatory plans are selected to hit these targets. Thus, essentially a target is the sound pattern that the talker is trying to articulate. According to current theory, auditory targets are integrated with speech motor plans via a dorsal sensorimotor[1] processing stream (Guenther, 2006; Hickok, 2012, 2014; Hickok et al., 2011; Indefrey and Levelt, 2004; Tourville et al., 2008). Various sources of evidence confirm that intact auditory speech representations are necessary for normal speech production. These include articulatory decline in adult-onset deafness (Waldstein, 1990), disruption of speech output by delayed auditory feedback (Stuart et al., 2002; Yates, 1963), and compensation for altered auditory feedback (Burnett et al., 1998; Purcell and Munhall, 2006).

Here we consider the role of *visual speech* in *production* and the neural circuits involved. Existing evidence shows that visual speech indeed plays such a role. A classic study (Reisberg et al., 1987) – one that, in fact, is frequently cited to support claims that visual speech increases auditory intelligibility – suggests that audiovisual speech facilitates production. In this study, participants were asked to *shadow* (listen to and immediately repeat word-by-word) spoken passages that were easy to hear but hard to understand – specifically, passages were spoken in a recently acquired second language, spoken in accented English, or drawn from semantically and syntactically complex content. The dependent variable was the tracking (speech production) rate in words per minute, and this rate significantly increased when spoken passages were accompanied by concurrent visual speech.

[1]Throughout the manuscript, we will use the terms 'motor' or 'senorimotor' to refer to operations that participate in generating speech. These may include sensory systems (the targets of motor control) and cognitive operations such as attention, working memory, sequencing, etc., that are necessary for normal speech production

Circumstantial evidence for the claim that visual speech supports production can be drawn from recent neurophysiological research indicating that visual and audiovisual speech activate the speech motor system (Callan et al., 2003; Hasson et al., 2007; Ojanen et al., 2005; Okada and Hickok, 2009; Skipper et al., 2007; Watkins et al., 2003). Although again this evidence is often interpreted to support a role for the motor system in visual or multimodal speech *perception* (Möttönen and Watkins, 2012; Schwartz et al., 2012), the reverse relation – visual speech supports *production* – is equally plausible. This notion motivated some of our own research examining the effects of visual speech on production. Upon observing that visual and auditory speech perception activate the speech motor system (Fridriksson et al., 2008; Rorden et al., 2008), we hypothesized that perceptual training with audiovisual speech would improve the speech output of nonfluent aphasics. Indeed, when patients were trained on a word-picture matching task, significant improvement in subsequent picture naming was observed, but only when the training phase included audiovisual words (Fridriksson et al., 2009). We later observed a striking effect we termed "Speech Entrainment" (SE), in which shadowing of audiovisual speech allowed patients with nonfluent aphasia to increase their speech output by a factor of two or more (Fridriksson et al., 2012). This effect was not observed for shadowing of unimodal auditory or visual speech, which suggests the following: motor commands for speech at some level of the motor hierarchy are relatively intact in some cases of nonfluent aphasia, and visual speech when combined with auditory speech provides crucial information allowing access to these motor commands.

As such, current evidence points to the conclusion that visual speech plays a role in speech motor control, at least in some situations. To be specific, we assume that the noted behavioral increases in speech output during or following exposure to audiovisual speech reflect the activation of a complementary set of visual speech targets (i.e., the visual patterns a talker is trying to produce) that combine with auditory speech targets to facilitate speech motor control. The origin of this visual influence on speech motor control likely comes via the role of visual speech in early acquisition of speech production, as is the case for auditory speech (Doupe and Kuhl, 1999; Ejiri, 1998; Oller and Eilers, 1988; Werker and Tees, 1999; Wernicke, 1969; Westermann and Miranda, 2004). To wit, human infants have a remarkable perceptual acuity for visual speech – they can detect temporal asynchrony in audiovisual speech (Dodd, 1979; Lewkowicz, 2010), match and integrate auditory and visual aspects of speech (Kuhl and Meltzoff, 1982; Lewkowicz, 2000; Patterson and Werker, 1999, 2003; Rosenblum et al., 1997), and they also show perceptual narrowing to native-language speech sounds perceived visually (Pons et al., 2009) as is true in the auditory modality (Werker and Tees, 1984). Moreover, neonates imitate at least some visually-perceived mouth movements (Abravanel and DeYong, 1991; Abravanel and Sigafoos, 1984; Kugiumutzakis, 1999; Meltzoff and Moore, 1977, 1983, 1994), and imitation increases over the first year of age with mouth movements forming a greater proportion of copied actions at early ages (Masur, 2006; Uzgiris et al., 1984). Crucially, infants who look more at their talking mother's mouth at 6 months of age score higher on expressive language at 24 months of age (Young et al., 2009), and infants begin to shift their attention from eyes to mouth between 4 and 8 months of age (Lewkowicz and Hansen-Tift, 2012) when babbling first begins to resemble speech-like sounds (Jusczyk, 2000; Oller, 2000); furthermore, attention shifts back to the eyes at 12

mo (Lewkowicz and Hansen-Tift, 2012) *after* narrowing toward native language speech patterns has occurred in both perception (Jusczyk, 2000; Werker and Tees, 1984) and production (cf., Werker and Tees, 1999).

It is likely that exposure to visual speech during acquisition of speech production establishes the neural circuitry linking visually-perceived gestures to the speech motor system. The behavioral and neuropsychological data reviewed above suggest such circuitry remains active in adulthood, as is the case for auditory speech (Hickok and Poeppel, 2007). A straightforward hypothesis concerning the mechanism for sensorimotor integration of visual speech is that visual speech is first combined with auditory speech via multisensory integration (e.g., in the posterior superior temporal lobe; Beauchamp et al., 2004; Beauchamp et al., 2010; Bernstein and Liebenthal, 2014; Calvert et al., 2000), and the resulting integrated speech representation is then fed into the auditory-motor dorsal stream either directly or by feedback with the auditory system (Arnal et al., 2009; Bernstein and Liebenthal, 2014; Calvert et al., 1999; Calvert et al., 2000; Okada et al., 2013). Another possibility is that visual speech activates dorsal stream networks directly, i.e., converging with auditory speech information not in (multi-) sensory systems but in the auditory-motor system. Indeed, speech-reading (perceiving visual speech) activates both multimodal sensory speech regions in the posterior superior temporal lobe and a well-known sensorimotor integration region for speech (Spt) in left posterior Sylvian cortex (Okada and Hickok, 2009). A third possibility is that dedicated sensorimotor networks exist for visual speech, distinct from the auditory-motor pathway. Speech Entrainment provides indirect support for this position – namely, the addition of visual speech improves speech output in a population for which canonical auditory-motor integration networks are often extensively damaged (Fridriksson et al., 2014; Fridriksson et al., 2012).

In the current study, we attempted to disambiguate these possibilities by studying the organization of sensorimotor networks for auditory, visual, and audiovisual speech. Specifically, we used blood-oxygen-level-dependent (BOLD) fMRI to test whether covert repetition of visual or audiovisual speech: (1) increased activation in known auditory-motor circuits (relative to repetition of auditory-only speech), (2) activated sensorimotor pathways unique to visual speech, (3) both, or (4) neither. Covert repetition[2] is often employed to identify auditory-to-vocal-tract networks (Buchsbaum et al., 2001; Hickok et al., 2003; Okada and Hickok, 2006; Rauschecker et al., 2008; Wildgruber et al., 2001). We adapted a typical covert repetition paradigm to test whether using visual (V) or audiovisual (AV) stimuli to cue repetition recruits different sensorimotor networks versus auditory-only (A) input. To explore the space of hypotheses listed above, i.e., (1), (2), (3), and (4), we identified regions demonstrating an increased rehearsal-related response for V or AV inputs relative to A. Briefly, the results demonstrate that using V or AV speech to cue covert reptition increases the extent of activation in canonical auditory-motor networks, and results in activation of additional speech motor brain regions (over and above A). Moreover, a sensorimotor region in the left posterior middle temporal gyrus activates preferentially during rehearsal and passive perception in the V and AV modalities, suggesting that this

---

[2]Covert production is preferred because overt production makes it difficult to distinguish activation to passive perception (i.e., via external feedback during overt production) versus activation to motor-rehearsal components of the task.

region may host visual speech targets for production. To the best of our knowledge, this is the first time this type of sensorimotor-speech design has been applied comprehensively to multiple input modalities within the same group of participants. The goal of this application was to provide an initial characterization of the sensorimotor networks for (audio-)visual speech, whereas putative auditory-motor networks have been characterized and studied extensively (Brown et al., 2008; Hickok et al., 2009; Price, 2014; Simmonds et al., 2014a; Simmonds et al., 2014b; Tourville et al., 2008).

2. Materials and Methods

2.1 Participants

Twenty (16 female) right-handed native English speakers between 20 and 30 years of age participated in the fMRI study. All volunteers had normal or corrected-to-normal vision, normal hearing by self-report, no known history of neurological disease, and no other contraindications for MRI. Informed consent was obtained from each participant in accordance with University of South Carolina Institutional Review Board guidelines. An additional nine native English speakers between 21 and 31 year of age participated in a follow-up behavioral study. One of these participants was left-handed. Volunteers for the follow-up study had normal or corrected-to-normal vision and normal hearing by self-report. Oral informed consent was obtained in accordance with UC Irvine Institutional Review Board guidelines.

2.2 Stimuli and Procedure

Forty-five digital video clips (3s duration, 30 fps) were produced featuring a single male talker shown from the neck up. In each clip, the talker produced a sequence of four consonant-vowel (CV) syllables drawn from a set of six visually distinguishable CVs: /ba/, /tha/, /va/, /bi/, /thi/, /vi/ (IPA: /ba/, /ða/, /va/, /bi/, /ði/, /vi/). The consonant was voiced for each individual CV. The CVs were articulated as a continuous sequence with the onset of each component syllable timed to a visual metronome at 2 Hz. Each of the six CVs appeared exactly 30 times across all 45 clips and any given CV was never repeated in a sequence. Otherwise, the ordinal position of each CV within a sequence was selected at random (see Supplementary Methods for more details concerning stimulus construction).

Syllable sequences were presented to participants in each of three modalities (Figure 1A): auditory-only (A), visual-only (V) and audiovisual (AV). In the A modality, clips consisted of a still frame of the talker's face paired with auditory recordings of CV syllable sequences. In the V modality, videos of the talker producing syllable sequences were presented without sound. In the AV modality, videos of the talker producing syllable sequences were presented along with the concurrent auditory speech signal. In addition, syllable sequences were presented in three experimental conditions: perception followed by covert rehearsal (P +RHRS), perception followed by rest (P+REST), and continuous perception (CP). A single trial in each condition comprised a 10s period (Figure 1B): a visual cue indicating the condition (1.5s) followed by a blank gray screen (uniformly jittered duration, 0.5–2s), stimulation (6s), and then a black fixation "X" on the gray background (remainder of 10s). Only the 6s stimulation period varied by condition. In the P+RHRS condition, participants

were asked to "perceive" (watch, listen to, or both) syllable sequences (3s) and then covertly rehearse (single repetition) the perceived sequence in the period immediately following (3s). In the P+REST condition, participants were asked to perceive syllable sequences (3s) followed immediately by a period of rest (3s without covert articulation). In the CP condition, participants were asked to perceive a syllable sequence that was presented twice so as to fill the entire stimulation period (6s). There were also rest (baseline) trials in which the black fixation "X" was presented for the entire 10s. Functional imaging runs consisted of 40 trials, 10 from each condition and 10 rest trials. Runs were blocked by modality – that is, of nine functional runs, there were three A runs, three V runs, and three AV runs, presented in pseudo-random order (the same modality was never repeated more than once and each participant encountered a different presentation order). Within each run, trials from each condition were presented in pseudo-random order (same condition never repeated more than once and rest trials were never repeated; see Supplementary Methods). Stimulus delivery and timing were controlled using the Psychtoolbox-3 (https://www.psychtoolbox.org/) implemented in Matlab (Mathworks Inc., USA). Videos were presented centrally and subtended 6.9°×5.0° of visual angle. The face and mouth of the talker subtended 3.0° and 1.5° of visual angle, respectively.

2.3 Follow-up Behavioral Study

In order to assess the difficulty of the P+RHRS task across presentation modalities (A, V, AV), a second group of participants was asked to complete an *overt* repetition task outside the scanner. The stimuli were the same as those used in the fMRI experiment. Stimulus presentation was carried out on a laptop computer running Ubuntu Linux v12.04. Stimulus delivery and timing were controlled using the Psychtoolbox-3 (https://www.psychtoolbox.org/) implemented in Matlab (Mathworks Inc., USA). Sounds were presented over Sennheiser HD 280 pro headphones. Participants were seated inside a double-walled, sound-attenuated chamber.

Participants were made explicitly aware of the stimulus set (/ba/, /tha/, /va/, /bi/, /thi/, /vi/) prior to the experiment. On each trial, a syllable sequence was presented (3s) followed by a fixation cross indicating that the participant should begin to repeat the sequence once out loud. Participants were asked to match the presentation rate of the syllable sequence (2 Hz) but the response period was unconstrained. Oral responses were recorded using the internal microphone of the stimulus presentation computer. After completing the oral response, the participant pressed the 'Return' key. At this time four boxes appeared on the screen, each of which was labeled using the true syllable (in the correct position) from the target sequence presented on that trial. The participant was instructed to select the syllables that were repeated in the correct position (i.e., correct his/her own response). Prior to response correction, the participant was instructed to play back the response over the headphones by pressing the 'p' key. Following response correction, the next trial was then initiated by pressing the 'Return' key. After the experiment was completed, the verbal utterance from each trial was played back and scored/checked manually by an experimenter (JHV).

Participants completed six practice trials, two in each presentation modality (A, V, AV), followed by 30 experimental trials (10 per modality). Trials were blocked by presentation

modality with order counterbalanced across participants. The particular syllable sequences presented to each participant were drawn at random from the set of 45 sequences used in the fMRI experiment. Performance on the task was quantified by three measures. First, we scored the mean number of syllables (per trial) correctly repeated *in the correct position* (N_{cor}) of the syllable sequence. Second, we scored the mean number of syllables (per trial) correctly repeated *regardless of position* (N_{any}) in the syllable sequence. Finally, we scored the mean number of utterances completed per trial, regardless of correctness, within three seconds from the onset of the fixation cross (N_{3s}); three seconds was chosen to match the duration of the response period during the fMRI experiment. These measures were tabulated separately for each modality, and separate repeated measures ANOVAs (Greenhouse-Geisser corrected) were conducted for each measure including a single factor (modality) with three levels (A, V, AV). Post-hoc pairwise comparisons were carried out using the paired-samples t-test with Bonferroni correction for multiple comparisons (threshold p-value 0.0167).

2.4 Scanning Parameters

MR images were obtained on a Siemens 3T fitted with a 12-channel head coil and an audio-visual presentation system. We collected a total of 1872 echo planar imaging (EPI) volumes per participant over 9 runs using single pulse Gradient Echo EPI (matrix = 104×104, repetition time [TR] = 2s, echo time [TE] = 30ms, size = $2 \times 2 \times 3.75$ mm, flip angle = 76). Thirty-one sequentially acquired axial slices provided whole brain coverage. After the functional scans, a high-resolution T1 anatomical image was acquired in the sagittal plane (1 mm^3).

2.5 Imaging Analysis –fMRI

Preprocessing of the data was performed using AFNI software (http://afni.nimh.nih.gov/afni). For each run, slice timing correction was performed followed by realignment (motion correction) and coregistration of the EPI images to the high resolution anatomical image in a single interpolation step. Functional data were then warped to a study-specific template (MNI152 space) created with Advanced Normalization Tools (ANTS; see Supplementary Methods). Finally, images were spatially smoothed with an isotropic 6-mm full-width half-maximum (FWHM) Gaussian kernel and each run was scaled to have a mean of 100 across time at each voxel.

First level regression analysis (AFNI 3dREMLfit) was performed in individual participants. The hemodynamic response function (HRF) for events from each cell of the design was estimated using a cubic spline (CSPLIN) function expansion with 8 parameters modeling the response from 2 to 16s after stimulation onset (spacing = 1 TR). The HRF was assumed to start (0s post-stimulation) and end (18s post-stimulation) at zero. The amplitude of the response was calculated by averaging the HRF values from 6–10s post-stimulation. A total of 72 regressors were used to model the HRF from each of the 9 event types in the experiment: A-P+RHRS, A-P+REST, A-CP, V-P+RHRS, V-P+REST, V-CP, AV-P+RHRS, AV-P+REST, AV-CP.

A second-level mixed effects analysis (AFNI 3dMEMA (Chen et al., 2012)) was performed on the HRF amplitude estimates from each participant, treating 'participant' as a random effect. This procedure is similar to a standard group-level t-test but also takes into account the level of intra-participant variation by accepting t-scores from each individual participant analysis. Statistical parametric maps (t-statistics) were created for each contrast of interest. Active voxels were defined as those for which t-statistics exceeded the $p < 0.005$ level with a cluster extent threshold of 173 voxels. This cluster threshold was determined by Monte Carlo simulation (AFNI 3dClustSim) to hold the family-wise error rate (FWER) less than 0.05 (i.e., corrected for multiple comparisons). Estimates of smoothness in the data were drawn from the residual error time series for each participant after first-level analysis (AFNI 3dFWHMx). These estimates were averaged across participants separately in each voxel dimension for input to 3dClustSim. Simulations were restricted to in-brain voxels.

We performed two group-level contrasts to identify different components of speech-related sensorimotor brain networks. The first, which we term 'perception only' (PO), tested for activation greater in the CP condition than baseline (CP > Rest) and was intended to identify brain regions involved in the sensory phase of the perceive+rehearse task. The second, which we term 'rehearsal minus perception' (RmP), tested for activation greater in the P+RHRS condition than the P+REST condition (P+RHRS > P+REST). This contrast factored out activation related to *passive perception* and was thus intended to identify brain regions involved in the (covert) rehearsal phase of the perceive+rehearse task. Finally, sensorimotor brain regions (i.e., those involved in both phases of the task) were identified by performing a conjunction (SM-CONJ) of PO and RmP contrasts (CP > rest ∩ P+RHRS > P+REST). The conjunction analysis was performed by constructing minimum *t*-maps (e.g., minimum T score from [PO, RmP] at each voxel) thresholded at $p < 0.005$ with a cluster extent threshold of 173 voxels (FWER < 0.05, as for individual condition maps). PO, RmP, and SM-CONJ were calculated separately for each modality to form a total of 9 group-level SPMs. We tested directly for differences in motor activation across input modalities by performing two interaction contrasts: VvsA-RmP (V-RmP > A-RmP) and AVvsA-RmP (AV-RmP > A-RmP). These interaction contrasts were masked inclusively with V-RmP and AV-RmP, respectively, in order to exclude voxels that reached significance due to differential deactivation. As such, the masked interaction contrasts revealed voxels that (a) activated significantly during the rehearsal phase of the task, and (b) demonstrated greater rehearsal-related activation when the input stimulus included visual speech.

Activations were visualized on the Conte69 atlas in MNI152 space in CARET v5.65 (http://brainvis.wustl.edu/wiki/index.php/Caret:Download), or on the study-specific template in MNI152 space in AFNI or MRIcron (http://www.mccauslandcenter.sc.edu/mricro/mricron/). Displayed group-average time-course plots were formed by taking the average of individual participant HRF regression parameters at each time point and performing cubic spline interpolation with a 0.1s time step. In the plots, time zero refers to the onset of the 6s stimulation period (Figure 1B).

3. Results

3.1 fMRI hypotheses (recap)

We hypothesized the following with respect to our multimodal covert rehearsal task: (1) Additional sensorimotor brain regions (i.e., outside canonical auditory-motor networks as assessed in the A modality) will be recruited when the input modality is V, AV, or both; (2) Additional rehearsal activation will be observed (either within canonical auditory-motor regions or in visual-specific regions) when the input modality is V, AV, or both (i.e., there will be rehearsal activation over and above that observed for auditory-only input). To assess (1), we simply observed differences in sensorimotor networks (SM-CONJ) across modalities qualitatively. To assess (2), we tested directly for differences in rehearsal (RmP) activation across modalities by performing two interaction contrasts: AVvsA-RmP and VvsA-RmP.

3.2 Sensorimotor speech networks for multiple input modalities

The SM-CONJ map in the A modality (Figure 2, top left) comprised significant clusters in canonical motor-speech brain regions including the left inferior frontal gyrus/frontal operculum (IFG), the left precentral gyrus (PreM), and bilateral supplementary motor area (SMA; preponderance of activation in the left hemisphere). In addition, there were significant clusters in the left posterior Sylvian region (Spt), right PreM, and the right cerebellum (cerebellar activations pictured in Figure 2, middle right). This network matches up quite well with previously identified auditory-motor integration networks for the vocal tract (Buchsbaum et al., 2001; Hickok et al., 2003; Isenberg et al., 2012; Okada and Hickok, 2006).

The SM-CONJ map in the V (Figure 2, top right) and AV (Figure 2, middle left) modalities included the same network of brain regions, but with increased extent of activation in several clusters (Table 1). Additionally, several new clusters emerged in the SM-CONJ for both V and AV, consistent with our hypothesis. Common to both maps, additional clusters were active in the left posterior superior temporal sulcus/middle temporal gyrus (STS/MTG), the right posterior STS, and the left Insula (all highlighted in colored boxes in Figure 2). Examination of activation time-courses (Figure 2, bottom) indicates that increased rehearsal activation in the V and AV modalities (relative to A) likely drove these additional clusters above threshold. Overall, sensorimotor integration of visual speech – whether V or AV – recruited the standard auditory-motor speech network more robustly as well as additional posterior superior temporal regions.

3.3 Explicit tests for increased rehearsal activation when sensory inputs contained visual speech

Figures 3 and 4 show significant clusters for the AVvsA-RmP and VvsA-RmP interaction contrasts, respectively (see Table 2 for MNI coordinates). For each cluster we have plotted the activation time-courses for 'passive perception' (PO) and 'rehearsal minus perception' (RmP) in each sensory modality (A, V, AV). Most of these clusters demonstrated a strong RmP response in the V and AV modalities but little or no RmP response in the A modality, and little or no PO response in any sensory modality. In some cases, e.g. the left IFG, there was a discernible RmP response in all three modalities, but the response was graded

(favoring V and AV). Only the left posterior MTG showed a strong PO response in addition to an RmP response, with greater activation to V and AV in both PO and RmP. Crucially, the interaction contrasts primarily identified speech-motor-related brain regions including the inferior frontal gyrus, precentral gyrus, central sulcus, insula, cerebellum, and caudate nucleus (cf., Eickhoff et al., 2009). Several of these regions responded only to V and/or AV while other regions responded to all three modalities but responded best in the presence of a visual speech signal.

3.4 Behavior

A significant main effect of modality was observed for each performance measure (N_{cor}: $F(1.82, 14.59) = 21.0$, $p < 0.001$; N_{any}: $F(1.56, 12.49) = 10.67$, $p = 0.003$; N_{3s}: $F(1.52, 12.20) = 10.14$, $p = 0.004$). For each measure, performance was significantly worse in the V modality. This was confirmed by paired-samples t-tests comparing A vs. V (N_{cor}: $t(8) = 4.99$, $p = 0.001$; N_{any}: $t(8) = 5.06$, $p = 0.001$; N_{3s}: $t(8) = 3.05$, $p = 0.016$) and AV vs. V (N_{cor}: $t(8) = 6.54$, $p < 0.001$; N_{any}: $t(8) = 4.00$, $p = 0.004$; N_{3s}: $t(8) = 3.66$, $p = 0.006$). There were no significant differences between the A and AV modalities although performance tended to be best in the AV modality. It should be noted that several participants performed at ceiling in the A and AV modalities for one of our performance measures (N_{3s}). Overall these results demonstrate that participants were significantly less accurate and also slower to respond in the V modality relative to A and AV. However, the observed effects were small and overall performance in each modality was quite good, indicating that participants were able to adequately perform the task with minimal training.

4. Discussion

In the current study we asked whether sensorimotor networks for speech differed depending on the sensory modality of the input stimulus. In particular, we conducted an fMRI experiment in which participants were asked to perceive and covertly repeat a sequence of consonant-vowel syllables presented in one of three sensory modalities: auditory (A), visual (V), or audiovisual (AV). We measured activation to both passive perception (PO) and rehearsal (RmP) of the input stimuli and we identified sensorimotor brain regions by testing for voxels that activated significantly to perception and rehearsal (SM-CONJ). We also tested for regions showing an increased rehearsal response when visual speech was included in the input (V or AV) relative to auditory-only input. Despite the fact that the motor act of rehearsing a set of syllables is *prima facie* the same when the set is cued with or without a visible talker, we hypothesized that inclusion of visual speech in the input would either augment the activation in known auditory-motor networks (via multisensory integration of AV inputs) or recruit additional sensorimotor regions. Three noteworthy results will be discussed at further length below. First, speech motor regions were more activated when the input stimulus included visual speech. Second, certain motor and sensorimotor regions were activated *only* when the input stimulus included visual speech. Third, regions that activated preferentially for V input also tended to activate well to AV input and vice versa. A behavioral experiment conducted outside the scanner demonstrated that modality differences in neural activity – particularly those between AV and A – could not be accounted for in terms of differences in task difficulty across the modalities.

4.1 Visual speech inputs increase speech motor activation during rehearsal

Two sources of evidence support the conclusion that visual speech inputs increase speech motor activation during rehearsal. The first concerns differences in the SM-CONJ maps for V and AV compared to A. Both V and AV inputs increased the extent of sensorimotor activation in canonical speech motor regions including the ventral premotor cortex. This "gain increase," while noticeable, was a relatively minor effect. A more striking effect was the emergence of an additional significant cluster in the left insula in the SM-CONJ maps for both V and AV (i.e., the cluster was not part of the SM-CONJ map for A). Examination of the activation time-course in this left insula region indicates that increased rehearsal-related activation in the V and AV modalities (relative to A) drove the effect.

The second and most significant source of evidence comes from direct contrasts of rehearsal-related activation for visual speech inputs (V, AV) versus auditory-only inputs (A). The AVvsA-RmP contrast revealed significant clusters in left rolandic cortex and bilateral precentral regions. The VvsA-RmP contrast revealed clusters in the left IFG, insula, caudate nucleus, and the right cerebellum. All of these clusters showed little or no activation during PO in any of the sensory modalities. In short, covert rehearsal of stimuli containing visual speech preferentially activated a network of speech motor brain regions, and these brain regions did not activate significantly above baseline during passive perception of stimuli containing visual speech.

These results concur with behavioral evidence from normal and aphasic individuals indicating that shadowing audiovisual speech leads to increased speech output compared to shadowing of auditory-only speech (Fridriksson et al., 2012; Reisberg et al., 1987). Specifically, we have shown here that rehearsal immediately following speech input leads to greater activation in speech motor regions when the input contains visual speech, with effects in primary motor regions for audiovisual speech in particular. However, there is one potential flaw in the argument that activation of motor regions reflects motor processing *per se*, viz. that differences in *perceptual* processing between the input modalities produced the observed differences in activation in motor brain areas. The RmP contrast, which subtracts out the activation to *passive perception* of the sensory inputs, is intended to identify a network of brain regions that perform various computations related to covert rehearsal: attention to sensory targets (potentially leading to increased activation of the targets themselves), maintenance of those targets in memory, integration of sensory information with the motor system, selection and activation of motor programs, and internal feedback control. It is therefore impossible, without a detailed analysis that is beyond the scope of the current study, to specify exactly which computation(s) a particular RmP cluster performs. Indeed *any* RmP cluster identified using the current methods might be involved in perception- rather than motor-related components of the covert rehearsal task. Crucially, several authors have suggested a role for motor speech brain regions in perceptual processing of visual speech (Callan et al., 2003; Hasson et al., 2007; Ojanen et al., 2005; Okada and Hickok, 2009; Skipper et al., 2007; Watkins et al., 2003). Hence, one could argue that modality differences (V/AV vs. A) in the rehearsal networks reported here – even in canonical 'motor' brain regions – were driven entirely by increased attention to the visual speech signal in service of perception (i.e., prior to or separate from performance of covert

motor acts). It is very unlikely that this is the case. Firstly, we have argued elsewhere that motor brain regions in fact do not contribute significantly to speech perception (for reviews see: Hickok, 2009; Venezia and Hickok, 2009; Venezia et al., 2012), including perception of visual speech (Matchin et al., 2014; for alternative views see: Callan et al., 2003; Hasson et al., 2007; Skipper et al., 2007). Secondly, examination of activation time-courses in the speech motor brain regions identified in the current study (Figs. 3–4) indicates that these regions did not respond above baseline during passive perception of visual or audiovisual speech. While it is quite likely that participants allocated less attention on passive perception trials, they were nonetheless instructed to pay attention to the stimulus, and previous work suggests that passive perception of visual speech should be sufficient to drive the relevant motor areas (Okada and Hickok, 2009; Sato et al., 2010; Turner et al., 2009; Watkins et al., 2003). Indeed, in the current study passive perception of visual and audiovisual speech strongly engaged speech motor regions in the left and right hemispheres (Supplementary Figure 1); these were simply not the same regions that showed interaction effects (VvsA, AVvsA) in terms of rehearsal-related activation.

4.2 A distinct sensorimotor pathway for visual speech

Several of the sensorimotor brain regions that responded more during rehearsal of V or AV inputs also responded during rehearsal of A inputs. These include the bilateral pSTS and left insula, ventral premotor cortex, and inferior parietal lobule. This set of regions responded *preferentially* to V, AV or both (relative to A). Other brain regions showing increased rehearsal-related activation following visual speech input responded *exclusively* to V, AV or both. Among these regions were the bilateral pre-central sulci and left central sulcus, caudate nucleus, IFG, and MTG. This finding supports the conclusion that visual speech representations have access to a distinct pathway to the motor system that, when engaged in conjunction with auditory-motor networks by an audiovisual stimulus, produces increased activation of speech motor programs. If so, the influence of visual speech on production cannot be reduced to secondary activation of canonical auditory-motor pathways (e.g., via activation of auditory-phonological targets that interface with the speech motor system; Calvert et al., 1999; Calvert et al., 1997; Calvert et al., 2000; Okada and Hickok, 2009; Okada et al., 2013).

Our results suggest that the left posterior MTG is a crucial node in the visual-to-motor speech pathway. This region was identified in the SM-CONJ for V and AV but not for A, and responded significantly more (in fact only responded) to RmP when the input stimulus was V or AV. Passive perceptual activation in the left pMTG was much greater for V and AV inputs as well (Figs. 3–4). The left pMTG figured prominently in a previous imaging study examining the effects of Speech Entrainment (SE) in nonfluent aphasics and normal participants (Fridriksson et al., 2012). In both participant groups, there was significantly greater activation in the left MTG for SE (audiovisual shadowing) compared to spontaneous speech production, and probabilistic fiber tracking based on DTI data in the normal participants indicated anatomical connections between the left pMTG and left inferior frontal speech regions via the arcuate fasciculus.

While we can only speculate on the precise role of the left pMTG in the visuomotor speech pathway, the large activations to passive perception of V and AV speech observed in this region suggest that it may play host to visual speech targets for production (i.e., high-level sensory representations of visual speech gestures). A number of previous studies have also found activation in the left pMTG during perception of visual or audiovisual speech (cf., Callan et al., 2003; Calvert and Campbell, 2003; Campbell et al., 2001; MacSweeney et al., 2001; Sekiyama et al., 2003). More recent work has shown that a particular region of the left pMTG/STS, dubbed the temporal visual speech area (TVSA; Bernstein et al., 2011), is selective for visual speech versus nonspeech facial gestures and also shows tuning to visual-speech phoneme categories (Bernstein et al., 2011; Bernstein and Liebenthal, 2014; De Winter et al., 2015; Files et al., 2013). The location of the TVSA is consistent with the left pMTG/STS activations observed in the current study. Importantly, the TVSA is a unimodal visual region that plays host to high-level representations of speech (Bernstein and Liebenthal, 2014), which is consistent with a role for the left pMTG in coding *visual* sensory targets for motor control.

4.3 Visual and audiovisual speech inputs engage a similar rehearsal network

Behavioral increases in speech output are observed when participants repeat audiovisual speech but not visual-only speech (Fridriksson et al., 2012; Reisberg et al., 1987). This is likely due to the fact that speechreading (i.e., of V alone) is perceptually demanding to the point that even the best speechreaders (with normal hearing) discern only 50–70% of the content from connected speech (Auer and Bernstein, 2007; MacLeod and Summerfield, 1987; Summerfield, 1992). As such, we may have expected to see differences in rehearsal-related activation depending on whether the input stimulus was AV or V. We did observe some such differences. However, when we specifically examined activation time-courses, it was generally the case that speech motor areas activated during repetition of AV inputs also tended to activate when repeating V inputs (Figs. 3–4). There are two possible reasons for this phenomenon. First, we used a closed stimulus set with CV syllables that were intended to be easily distinguishable in the V modality, such that speechreading performance would be much higher than that observed for connected speech (i.e., effects of perceptual difficulty on rehearsal, and thus modality differences on this basis, were nullified to some extent). Second, as observed above, there may be a distinct visual-to-motor pathway that integrates visual speech with the motor system, and this pathway should be similarly engaged by visual and audiovisual speech.

It is worth noting that, although the syllables presented in the current study were drawn from a closed set, the covert repetition task was slightly more difficult than in the V modality than in the AV or A modalities, as evidenced by the significantly poorer performance in the V modality observed in the follow-up behavioral study. This is particularly relevant with respect to certain regions that responded more during rehearsal of V inputs relative to AV or A. These include the left caudate nucleus, cingulate cortex, IFG, and insula. The caudate is part of the basal ganglia, a group of subcortical regions theorized to be crucial for sequencing and timing in speech production (Bohland et al., 2010; Fridriksson et al., 2005; Guenther, 2006; Lu et al., 2010a; Lu et al., 2010b; Pickett et al., 1998; Stahl et al., 2011), while the left IFG and cingulate cortex are crucial for conflict monitoring and resolving

among competing alternatives (Botvinick et al., 2004; Carter et al., 1998; January et al., 2009; Kerns et al., 2004; Novick et al., 2005; Novick et al., 2010). In addition, recent data implicate the insula in time processing, focal attention, and cognitive control (Chang et al., 2012; Kosillo and Smith, 2010; Menon and Uddin, 2010; Nelson et al., 2010). The role of these regions in higher-level cognitive operations may explain the increased activation observed in the V covert rehearsal task. The same cannot be argued for AV, which was the easiest version of the task, suggesting that differences between the common V-AV rehearsal network and the A rehearsal network were not driven by task difficulty.

4.4 Why is visual speech linked to the motor system?

There is a well-accepted answer to the question of why *auditory* information is linked to the motor system: speech sound representations are used to guide speech production. Development of the ability to speak constitutes the most intuitive evidence for this claim. In short, development of speech is a motor learning task that must take sensory speech as the input – at first from other users of the language and subsequently from self-generated babbling (Oller and Eilers, 1988). It has been suggested that a dorsal, auditory-motor processing stream functions to support language development, and that this stream continues to function into adulthood (Hickok and Poeppel, 2000; Hickok and Poeppel, 2004; Hickok and Poeppel, 2007). More recent models suggest that for adult speakers auditory input functions primarily to tune internal feedback circuits that engage stored speech-sound representations to guide online speech production in real time (Hickok, 2012).

We have asserted that an additional sensorimotor pathway functions to integrate visual speech information with the motor system. As suggested by developmental work reviewed in the Introduction, the neural circuitry for visuomotor integration of speech likely solidifies during early acquisition of speech production abilities, as is the case for the auditory-motor dorsal stream. While it seems rather uncontroversial to suggest that visual speech supports production during development, it is unclear what ecological function, if any, this visuomotor speech pathway continues to serve in healthy adults such as those tested in the current study. One possibility is that feedback from visual speech is used to tune internal vocal tract control circuits in a similar fashion to auditory speech. On the surface, this proposition is dubious because talkers do not typically perceive their own speech visually during articulation. However, it is possible that visual speech information from *other* talkers influences one's own production patterns. There is clear evidence in the auditory domain that features of ambient speech including voice pitch and vowel patterns are unintentionally (automatically) reproduced in listeners' own speech (Cooper and Lauritsen, 1974; Delvaux and Soquet, 2007; Kappes et al., 2009), an effect which some have interpreted to serve cooperative social functions (Pardo, 2006) in addition to shaping phonological systems in language communities (Delvaux and Soquet, 2007; Pardo, 2006). Similarly, adult humans automatically imitate visually-perceived facial expressions (as well as other postures and mannerisms), which has the effect of smoothing social interactions (Chartrand and Bargh, 1999). It is possible that adult humans automatically make use of (i.e., mimic) ambient visual speech patterns in service of such social functions. To be sure, the time-varying visual speech signal is a rich source of information from which a large proportion of acoustic speech patterns can be recovered (Jiang et al., 2002; Yehia et al., 1998; Yehia et al., 2002).

This suggests that visual speech is capable of driving complex changes in patterns of articulation over time.

Another possibility is that talkers control their visible speech in order for it to be seen. For instance, talkers adapt visual elements of speech produced in noise such that articulated movements are larger and more correlated with speech acoustics, resulting in an increased perceptual benefit from visible speech (Kim et al., 2011). These adaptations occur automatically in noise but are also part of explicit communicative strategies (Garnier et al., 2010; Hazan and Baker, 2011). Indeed, the increased audiovisual benefit for speech produced in noise is even larger when speech is produced in a face-to-face setting versus a non-visual control setting (Fitzpatrick et al., 2015). However, despite the feasibility of these suggestions, it remains to be fully explained why rehearsal of visual speech activates a *complementary* set of speech motor regions in healthy adults as observed here. A possibility is that complementary sets of instructions to the vocal tract articulators can be engaged by the visuomotor speech pathway.

5. Conclusions

In summary, we have demonstrated that covert rehearsal following perception of syllable sequences results in increased speech motor activation when the input sequence contains visual speech, despite the apparent similarity of the motor speech act itself. This increased activation is likely produced via recruitment of a visual-speech-specific network of sensorimotor brain regions. We presume this network functions to support speech motor control by providing a complementary set of visual speech targets that can be used in combination with auditory targets to guide production. This predicts that improvements in speech fluency will be observed when this visuomotor speech pathway is activated in conjunction with canonical auditory-motor speech pathways, a hypothesis in need of further testing. We have argued that the visuomotor speech stream is formed during development to facilitate acquisition of speech production. The visuomotor speech stream is capable of influencing speech output in neuropsychological patients and during some laboratory tasks, but whether and how this stream functions in healthy adults outside the laboratory remains to be more clearly elucidated.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Abbreviations

SE   speech entrainment

A   auditory

V	visual
AV	audiovisual
CV	consonant vowel
P+RHRS	perception followed by covert rehearsal
P+REST	perception followed by rest
CP	continuous perception
MNI	Montreal Neurological Institute
FWER	familywise error rate
PO	perception only
RmP	rehearsal minus perception
SM-CONJ	conjunction of PO and RmP
SPM	statistical parametric map
VvsA-RmP	greater RmP activity in V than A
AVvsV-RmP	greater RmP activity in AV than A
Spt	Sylvian parietotemporal
IFG	inferior frontal gyrus
PreM	premotor cortex
SMA	supplementary motor area
STS	superior temporal sulcus
MTG	middle temporal gyrus
DTI	diffusion tensor imaging
TVSA	temporal visual speech area

References

Abravanel E, DeYong NG. Does object modeling elicit imitative-like gestures from young infants? J Exp Child Psychol. 1991; 52:22–40. [PubMed: 1890381]

Abravanel E, Sigafoos AD. Exploring the presence of imitation during early infancy. Child development. 1984:381–392. [PubMed: 6723442]

Arnal LH, Morillon B, Kell CA, Giraud AL. Dual neural routing of visual facilitation in speech processing. J Neurosci. 2009; 29:13445–13453. [PubMed: 19864557]

Arnold P, Hill F. Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. British Journal of Psychology. 2001; 92:339–355.

Auer ET, Bernstein LE. Enhanced visual speech perception in individuals with early-onset hearing impairment. Journal of Speech, Language, and Hearing Research. 2007; 50:1157–1165.

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

Beauchamp MS, Nath AR, Pasalar S. fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. The Journal of Neuroscience. 2010; 30:2414–2417. [PubMed: 20164324]

Bernstein LE, Jiang J, Pantazis D, Lu ZL, Joshi A. Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. Hum Brain Mapp. 2011; 32:1660–1676. [PubMed: 20853377]

Bernstein LE, Liebenthal E. Neural pathways for visual speech perception. Frontiers in neuroscience. 2014:8. [PubMed: 24550770]

Bohland JW, Bullock D, Guenther FH. Neural representations and mechanisms for the performance of simple speech sequences. Journal of Cognitive Neuroscience. 2010; 22:1504–1529. [PubMed: 19583476]

Botvinick MM, Cohen JD, Carter CS. Conflict monitoring and anterior cingulate cortex: an update. Trends in Cognitive Sciences. 2004; 8:539–546. [PubMed: 15556023]

Brown S, Ngan E, Liotti M. A larynx area in the human motor cortex. Cerebral Cortex. 2008; 18:837–845. [PubMed: 17652461]

Buchsbaum BR, Hickok G, Humphries C. Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. Cognitive Science. 2001; 25:663–678.

Burnett TA, Freedland MB, Larson CR, Hain TC. Voice F0 responses to manipulations in pitch feedback. The Journal of the Acoustical Society of America. 1998; 103:3153–3161. [PubMed: 9637026]

Callan DE, Jones JA, Munhall K, Callan AM, Kroos C, Vatikiotis-Bateson E. Neural processes underlying perceptual enhancement by visual speech gestures. Neuroreport. 2003; 14:2213–2218. [PubMed: 14625450]

Calvert G, Campbell R. Reading speech from still and moving faces: the neural substrates of visible speech. Cognitive Neuroscience, Journal of. 2003; 15:57–70.

Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, David AS. Response amplification in sensory-specific cortices during crossmodal binding. Neuroreport. 1999; 10:2619. [PubMed: 10574380]

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS. Activation of auditory cortex during silent lipreading. Science. 1997; 276:593–596. [PubMed: 9110978]

Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Biology. 2000; 10:649–658. [PubMed: 10837246]

Campbell R, MacSweeney M, Surguladze S, Calvert G, McGuire P, Suckling J, Brammer MJ, David AS. Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). Cognitive Brain Research. 2001; 12:233–243. [PubMed: 11587893]

Carter CS, Braver TS, Barch DM, Botvinick MM, Noll D, Cohen JD. Anterior cingulate cortex, error detection, and the online monitoring of performance. Science. 1998; 280:747–749. [PubMed: 9563953]

Chang LJ, Yarkoni T, Khaw MW, Sanfey AG. Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. Cerebral Cortex. 2012:bhs065.

Chartrand TL, Bargh JA. The chameleon effect: The perception–behavior link and social interaction. Journal of personality and social psychology. 1999; 76:893. [PubMed: 10402679]

Chen G, Saad ZS, Nath AR, Beauchamp MS, Cox RW. FMRI group analysis combining effect estimates and their variances. Neuroimage. 2012; 60:747–765. [PubMed: 22245637]

Cooper WE, Lauritsen MR. Feature processing in the perception and production of speech. 1974

De Winter FL, Zhu Q, Van den Stock J, Nelissen K, Peeters R, de Gelder B, Vanduffel W, Vandenbulcke M. Lateralization for dynamic facial expressions in human superior temporal sulcus. Neuroimage. 2015; 106:340–352. [PubMed: 25463458]

Delvaux V, Soquet A. The influence of ambient speech on adult speech productions through unintentional imitation. PHONETICA-BASEL. 2007; 64:145.

Dodd B. Lip reading in infants: Attention to speech presented in-and out-of-synchrony. Cogn Psychol. 1979; 11:478–484. [PubMed: 487747]

Doupe AJ, Kuhl PK. Birdsong and human speech: common themes and mechanisms. Annu Rev Neurosci. 1999; 22:567–631. [PubMed: 10202549]

Eickhoff SB, Heim S, Zilles K, Amunts K. A systems perspective on the effective connectivity of overt speech production. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. 2009; 367:2399–2421.

Ejiri K. Relationship between rhythmic behavior and canonical babbling in infant vocal development. Phonetica. 1998; 55:226–237. [PubMed: 9933781]

Erber NP. Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech, Language, and Hearing Research. 1969; 12:423–425.

Files BT, Auer ET Jr, Bernstein LE. The visual mismatch negativity elicited with visual speech stimuli. Front Hum Neurosci. 2013:7. [PubMed: 23372547]

Fitzpatrick M, Kim J, Davis C. The effect of seeing the interlocutor on auditory and visual speech production in noise. Speech Communication. 2015; 74:37–51.

Fridriksson J, Baker JM, Whiteside J, Eoute D, Moser D, Vesselinov R, Rorden C. Treating visual speech perception to improve speech production in nonfluent aphasia. Stroke. 2009; 40:853–858. [PubMed: 19164782]

Fridriksson J, Fillmore P, Guo D, Rorden C. Chronic Broca's Aphasia Is Caused by Damage to Broca's and Wernicke's Areas. Cerebral Cortex. 2014:bhu152.

Fridriksson J, Hubbard HI, Hudspeth SG, Holland AL, Bonilha L, Fromm D, Rorden C. Speech entrainment enables patients with Broca's aphasia to produce fluent speech. Brain. 2012; 135:3815–3829. [PubMed: 23250889]

Fridriksson J, Moss J, Davis B, Baylis GC, Bonilha L, Rorden C. Motor speech perception modulates the cortical language areas. Neuroimage. 2008; 41:605–613. [PubMed: 18396063]

Fridriksson J, Ryalls J, Rorden C, Morgan PS, George MS, Baylis GC. Brain damage and cortical compensation in foreign accent syndrome. Neurocase. 2005; 11:319–324. [PubMed: 16251133]

Garnier M, Henrich N, Dubois D. Influence of sound immersion and communicative interaction on the Lombard effect. Journal of Speech, Language, and Hearing Research. 2010; 53:588–608.

Guenther FH. Cortical interactions underlying the production of speech sounds. J Commun Disord. 2006; 39:350–365. [PubMed: 16887139]

Hardison DM. Acquisition of second-language speech: Effects of visual cues, context, and talker variability. Applied Psycholinguistics. 2003; 24:495–522.

Hasson U, Skipper JI, Nusbaum HC, Small SL. Abstract coding of audiovisual speech: beyond sensory representation. Neuron. 2007; 56:1116–1126. [PubMed: 18093531]

Hazan V, Baker R. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditionsa). The Journal of the Acoustical Society of America. 2011; 130:2139–2152. [PubMed: 21973368]

Hickok G. Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans. Journal of Cognitive Neuroscience. 2009; 21:1229–1243. [PubMed: 19199415]

Hickok G. Computational neuroanatomy of speech production. Nature Reviews Neuroscience. 2012; 13:135–145.

Hickok G. Towards an integrated psycholinguistic, neurolinguistic, sensorimotor framework for speech production. Language, Cognition and Neuroscience. 2014; 29:52–59.

Hickok G, Buchsbaum B, Humphries C, Muftuler T. Auditory–motor interaction revealed by fMRI: speech, music, and working memory in area Spt. Cognitive Neuroscience, Journal of. 2003; 15:673–682.

Hickok G, Houde J, Rong F. Sensorimotor integration in speech processing: computational basis and neural organization. Neuron. 2011; 69:407–422. [PubMed: 21315253]

Hickok G, Okada K, Serences JT. Area Spt in the human planum temporale supports sensory-motor integration for speech processing. J Neurophysiol. 2009; 101:2725–2732. [PubMed: 19225172]

Hickok G, Poeppel D. Towards a functional neuroanatomy of speech perception. Trends in Cognitive Sciences. 2000; 4:131–138. [PubMed: 10740277]

Hickok G, Poeppel D. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition. 2004; 92:67–99. [PubMed: 15037127]

Hickok G, Poeppel D. The cortical organization of speech processing. Nat Rev Neurosci. 2007; 8:393–402. [PubMed: 17431404]

Indefrey P, Levelt WJ. The spatial and temporal signatures of word production components. Cognition. 2004; 92:101–144. [PubMed: 15037128]

Isenberg AL, Vaden KI, Saberi K, Muftuler LT, Hickok G. Functionally distinct regions for spatial processing and sensory motor integration in the planum temporale. Hum Brain Mapp. 2012; 33:2453–2463. [PubMed: 21932266]

January D, Trueswell JC, Thompson-Schill SL. Co-localization of stroop and syntactic ambiguity resolution in Broca's area: implications for the neural basis of sentence processing. Journal of Cognitive Neuroscience. 2009; 21:2434–2444. [PubMed: 19199402]

Jiang J, Alwan A, Keating PA, Auer ET, Bernstein LE. On the relationship between face movements, tongue movements, and speech acoustics. EURASIP Journal on Applied Signal Processing. 2002; 11:1174–1188.

Jusczyk, PW. The discovery of spoken language. MIT press; 2000.

Kappes J, Baumgaertner A, Peschke C, Ziegler W. Unintended imitation in nonword repetition. Brain and Language. 2009; 111:140–151. [PubMed: 19811813]

Kerns JG, Cohen JD, MacDonald AW, Cho RY, Stenger VA, Carter CS. Anterior cingulate conflict monitoring and adjustments in control. Science. 2004; 303:1023–1026. [PubMed: 14963333]

Kim J, Sironic A, Davis C. Hearing speech in noise: Seeing a loud talker is better. Perception-London. 2011; 40:853. [PubMed: 22128557]

Kosillo á, Smith A. The role of the human anterior insular cortex in time processing. Brain Structure and Function. 2010; 214:623–628. [PubMed: 20512365]

Kugiumutzakis G. Genesis and development of early infant mimesis to facial and vocal models. 1999

Kuhl, PK.; Meltzoff, AN. The bimodal perception of speech in infancy. American Association for the Advancement of Science; 1982.

Lewkowicz DJ. Infants' perception of the audible, visible, and bimodal attributes of multimodal syllables. Child development. 2000; 71:1241–1257. [PubMed: 11108094]

Lewkowicz DJ. Infant perception of audio-visual speech synchrony. Developmental psychology. 2010; 46:66. [PubMed: 20053007]

Lewkowicz DJ, Hansen-Tift AM. Infants deploy selective attention to the mouth of a talking face when learning speech. Proceedings of the National Academy of Sciences. 2012; 109:1431–1436.

Lu C, Chen C, Ning N, Ding G, Guo T, Peng D, Yang Y, Li K, Lin C. The neural substrates for atypical planning and execution of word production in stuttering. Experimental neurology. 2010a; 221:146–156. [PubMed: 19879262]

Lu C, Peng D, Chen C, Ning N, Ding G, Li K, Yang Y, Lin C. Altered effective connectivity and anomalous anatomy in the basal ganglia-thalamocortical circuit of stuttering speakers. Cortex. 2010b; 46:49–67. [PubMed: 19375076]

MacLeod A, Summerfield Q. Quantifying the contribution of vision to speech perception in noise. British Journal of Audiology. 1987; 21:131–141. [PubMed: 3594015]

MacSweeney M, Campbell R, Calvert GA, McGuire PK, David AS, Suckling J, Andrew C, Woll B, Brammer MJ. Dispersed activation in the left temporal cortex for speech-reading in congenitally deaf people. Proceedings of the Royal Society of London B: Biological Sciences. 2001; 268:451–457.

Masur EF. Vocal and action imitation by infants and toddlers during dyadic interactions. Imitation and the social mind: Autism and typical development. 2006

Matchin W, Groulx K, Hickok G. Audiovisual speech integration does not rely on the motor system: Evidence from articulatory suppression, the mcgurk effect, and fmri. Journal of Cognitive Neuroscience. 2014; 26:606–620. [PubMed: 24236768]

McCormick B. Audio-visual discrimination of speech*. Clinical Otolaryngology & Allied Sciences. 1979; 4:355–361. [PubMed: 487634]

McGurk H, MacDonald J. Hearing lips and seeing voices. 1976

Meltzoff AN, Moore MK. Imitation of facial and manual gestures by human neonates. Science. 1977; 198:75–78. [PubMed: 17741897]

Meltzoff AN, Moore MK. Newborn infants imitate adult facial gestures. Child development. 1983:702–709. [PubMed: 6851717]

Meltzoff AN, Moore MK. Imitation, memory, and the representation of persons. Infant Behavior and Development. 1994; 17:83–99. [PubMed: 25147416]

Menon V, Uddin LQ. Saliency, switching, attention and control: a network model of insula function. Brain Structure and Function. 2010; 214:655–667. [PubMed: 20512370]

Möttönen R, Watkins KE. Using TMS to study the role of the articulatory motor system in speech perception. Aphasiology. 2012; 26:1103–1118. [PubMed: 22942513]

Neely KK. Effect of visual factors on the intelligibility of speech. The Journal of the Acoustical Society of America. 1956; 28:1275–1277.

Nelson SM, Dosenbach NU, Cohen AL, Wheeler ME, Schlaggar BL, Petersen SE. Role of the anterior insula in task-level control and focal attention. Brain Structure and Function. 2010; 214:669–680. [PubMed: 20512372]

Novick JM, Trueswell JC, Thompson-Schill SL. Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. Cognitive, Affective, & Behavioral Neuroscience. 2005; 5:263–281.

Novick JM, Trueswell JC, Thompson-Schill SL. Broca's area and language processing: Evidence for the cognitive control connection. Language and Linguistics Compass. 2010; 4:906–924.

Ojanen V, Mottonen R, Pekkola J, Jaaskelainen IP, Joensuu R, Autti T, Sams M. Processing of audiovisual speech in Broca's area. Neuroimage. 2005; 25:333–338. [PubMed: 15784412]

Okada K, Hickok G. Left posterior auditory-related cortices participate both in speech perception and speech production: Neural overlap revealed by fMRI. Brain and Language. 2006; 98:112–117. [PubMed: 16716388]

Okada K, Hickok G. Two cortical mechanisms support the integration of visual and auditory speech: a hypothesis and preliminary data. Neurosci Lett. 2009; 452:219–223. [PubMed: 19348727]

Okada K, Venezia JH, Matchin W, Saberi K, Hickok G. An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex. PLoS ONE. 2013; 8:e68959. [PubMed: 23805332]

Oller, DK. The emergence of the speech capacity. Psychology Press; 2000.

Oller DK, Eilers RE. The role of audition in infant babbling. Child development. 1988:441–449. [PubMed: 3359864]

Pardo JS. On phonetic convergence during conversational interaction. The Journal of the Acoustical Society of America. 2006; 119:2382–2393. [PubMed: 16642851]

Patterson ML, Werker JF. Matching phonetic information in lips and voice is robust in 4.5-month-old infants. Infant Behavior and Development. 1999; 22:237–247.

Patterson ML, Werker JF. Two-month-old infants match phonetic information in lips and voice. Developmental Science. 2003; 6:191–196.

Pickett ER, Kuniholm E, Protopapas A, Friedman J, Lieberman P. Selective speech motor, syntax and cognitive deficits associated with bilateral damage to the putamen and the head of the caudate nucleus: a case study. Neuropsychologia. 1998; 36:173–188. [PubMed: 9539237]

Pons F, Lewkowicz DJ, Soto-Faraco S, Sebastián-Gallés N. Narrowing of intersensory speech perception in infancy. Proceedings of the National Academy of Sciences. 2009; 106:10598–10602.

Price CJ. Sensory-to-motor integration during auditory repetition: A combined fMRI and lesion study. 2014

Purcell DW, Munhall KG. Compensation following real-time manipulation of formants in isolated vowels. The Journal of the Acoustical Society of America. 2006; 119:2288–2297. [PubMed: 16642842]

Rauschecker AM, Pringle A, Watkins KE. Changes in neural activity associated with learning to articulate novel auditory pseudowords by covert repetition. Hum Brain Mapp. 2008; 29:1231–1242. [PubMed: 17948887]

Reisberg D, Mclean J, Goldfield A. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. 1987

Rorden C, Davis B, George MS, Borckardt J, Fridriksson J. Broca's area is crucial for visual discrimination of speech but not non-speech oral movements. Brain stimulation. 2008; 1:383. [PubMed: 19421338]

Rosenblum LD, Schmuckler MA, Johnson JA. The McGurk effect in infants. Perception & Psychophysics. 1997; 59:347–357. [PubMed: 9136265]

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. Cerebral Cortex. 2007; 17:1147–1153. [PubMed: 16785256]

Sato M, Buccino G, Gentilucci M, Cattaneo L. On the tip of the tongue: modulation of the primary motor cortex during audiovisual speech perception. Speech Communication. 2010; 52:533–541.

Schwartz JL, Basirat A, Ménard L, Sato M. The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. Journal of Neurolinguistics. 2012; 25:336–354.

Sekiyama K, Kanno I, Miura S, Sugita Y. Auditory-visual speech perception examined by fMRI and PET. Neurosci Res. 2003; 47:277–287. [PubMed: 14568109]

Simmonds AJ, Leech R, Collins C, Redjep O, Wise RJ. Sensory-Motor Integration during Speech Production Localizes to Both Left and Right Plana Temporale. The Journal of Neuroscience. 2014a; 34:12963–12972. [PubMed: 25253845]

Simmonds AJ, Wise RJ, Collins C, Redjep O, Sharp DJ, Iverson P, Leech R. Parallel systems in the control of speech. Hum Brain Mapp. 2014b; 35:1930–1943. [PubMed: 23723184]

Skipper JI, van Wassenhove V, Nusbaum HC, Small SL. Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. Cereb Cortex. 2007; 17:2387–2399. [PubMed: 17218482]

Stahl B, Kotz SA, Henseler I, Turner R, Geyer S. Rhythm in disguise: why singing may not hold the key to recovery from aphasia. Brain. 2011:awr240.

Stuart A, Kalinowski J, Rastatter MP, Lynch K. Effect of delayed auditory feedback on normal speakers at two speech rates. The Journal of the Acoustical Society of America. 2002; 111:2237–2241. [PubMed: 12051443]

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

Summerfield Q. Lipreading and audio-visual speech perception. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 1992; 335:71–78. [PubMed: 1348140]

Tourville JA, Reilly KJ, Guenther FH. Neural mechanisms underlying auditory feedback control of speech. Neuroimage. 2008; 39:1429–1443. [PubMed: 18035557]

Turner TH, Fridriksson J, Baker J, Eoute D Jr, Bonilha L, Rorden C. Obligatory Broca's area modulation associated with passive speech perception. Neuroreport. 2009; 20:492. [PubMed: 19238107]

Uzgiris IC, Vasek ME, Benson JB. A longitudinal study of matching activity in mother-infant interaction. Infant Behavior and Development. 1984; 7:371.

Venezia JH, Hickok G. Mirror Neurons, the Motor System and Language: From the Motor Theory to Embodied Cognition and Beyond. Language and Linguistics Compass. 2009; 3:1403–1416.

Venezia JH, Saberi K, Chubb C, Hickok G. Response bias modulates the speech motor system during syllable discrimination. Frontiers in Psychology. 2012:3. [PubMed: 22291676]

Waldstein RS. Effects of postlingual deafness on speech production: implications for the role of auditory feedback. The Journal of the Acoustical Society of America. 1990; 88:2099–2114. [PubMed: 2269726]

Watkins KE, Strafella AP, Paus T. Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia. 2003; 41:989–994. [PubMed: 12667534]

Werker JF, Tees RC. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavior and Development. 1984; 7:49–63.

Werker JF, Tees RC. Influences on infant speech processing: Toward a new synthesis. Annual review of psychology. 1999; 50:509–535.

Wernicke, C. The symptom complex of aphasia. Proceedings of the Boston Colloquium for the Philosophy of Science; 1966/1968; Springer; 1969. p. 34-97.

Westermann G, Miranda ER. A new model of sensorimotor coupling in the development of speech. Brain and Language. 2004; 89:393–400. [PubMed: 15068923]

Wildgruber D, Ackermann H, Grodd W. Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: effects of syllable repetition rate evaluated by fMRI. Neuroimage. 2001; 13:101–109. [PubMed: 11133313]

Yates AJ. Delayed auditory feedback. Psychological Bulletin. 1963; 60:213. [PubMed: 14002534]

Yehia H, Rubin P, Vatikiotis-Bateson E. Quantitative association of vocal-tract and facial behavior. Speech Communication. 1998; 26:23–43.

Yehia HC, Kuratate T, Vatikiotis-Bateson E. Linking facial animation, head motion and speech acoustics. Journal of Phonetics. 2002; 30:555–568.

Young GS, Merin N, Rogers SJ, Ozonoff S. Gaze behavior and affect at 6 months: predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. Developmental Science. 2009; 12:798–814. [PubMed: 19702771]

Highlights

- Rehearsal cued by visual (vs. auditory) speech increases activation of speech motor system

- Rehearsal cued by visual speech activates regions outside auditory-motor network

- Results support existence of a visuomotor speech pathway

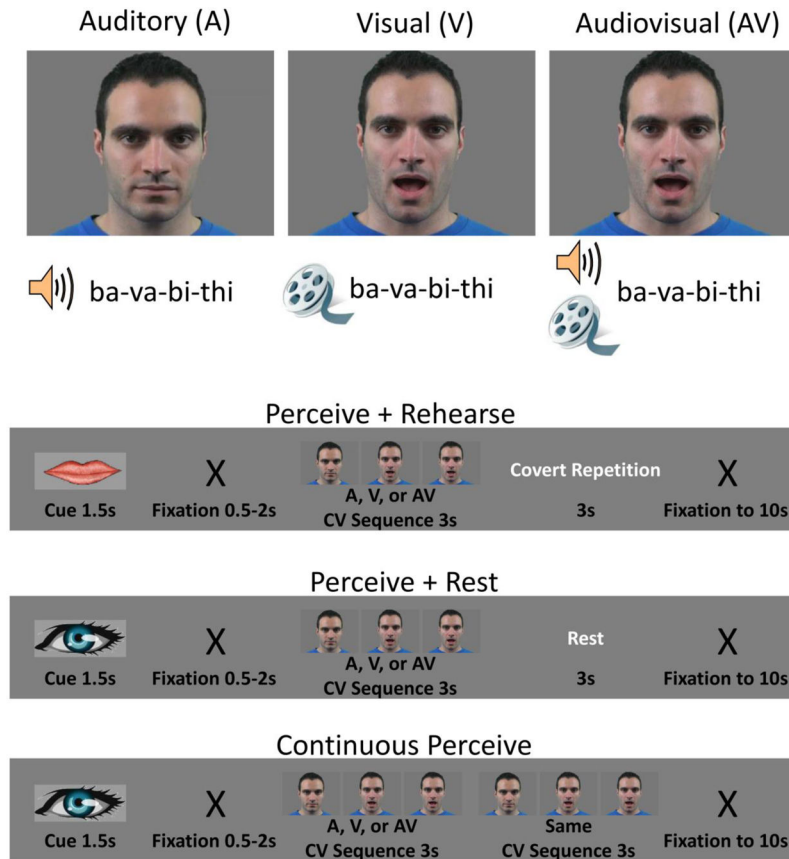- Results explain a puzzling effect from neuropsychology known as speech entrainment

Figure 1. Design schematic of multimodal sensorimotor speech task

Input modality (top) was crossed with condition (bottom). Each imaging run contained 30 trials from a given input modality – A, V or AV – and 10 rest trials (not pictured). Of the 30 trials, 10 each were perceive+rehearse (P+RHRS), perceive+rest (P+REST), and continuous perceive (CP). The P+RHRS trials were cued by an image of lips at the onset of the trial, and the P+REST and CP trials were cued by an image of an eye at the onset of the trial. Trial structures are shown for each condition. White text indicates what participants were actually doing (i.e., was not shown explicitly on the screen). Stimuli were 3s CV syllable sequences drawn from the set of visually distinguishable CVs /ba/, /bi/, /tha/, /thi/, /va/, /vi/. The CV sequence shown is just one possible example.
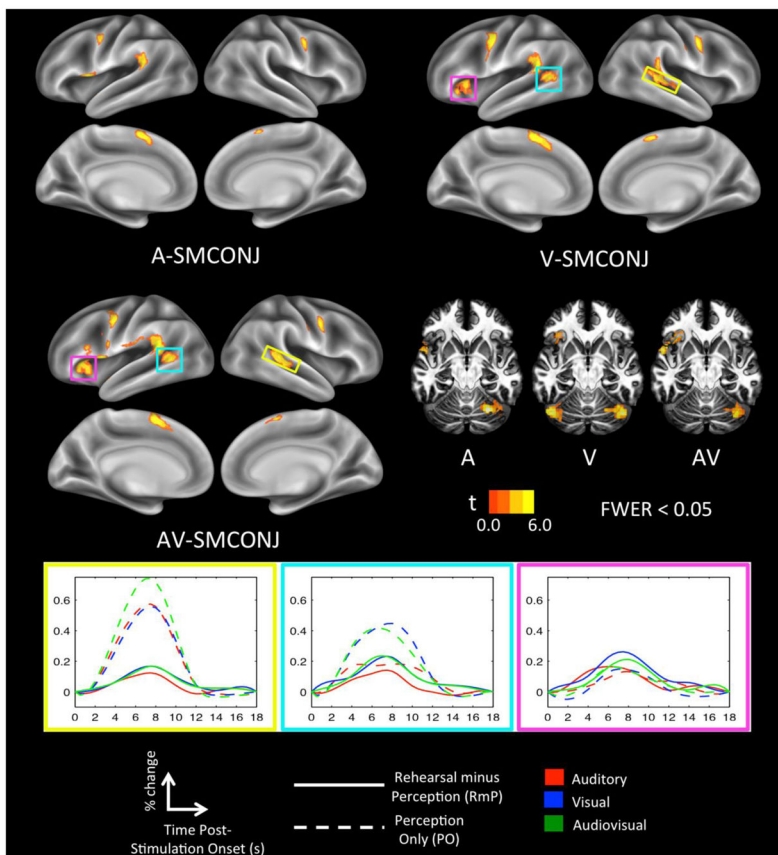
Figure 2. Sensorimotor conjunction maps
Sensorimotor brain regions were highlighted in each modality by taking the conjunction of 'perception only' (CP > baseline) and 'rehearsal minus perception' (P+RHRS > P+REST) contrasts. These regions are displayed on separate cortical surface renderings for each input modality (A, V, AV). Also plotted are show-through volume renderings for each modality with axial slices peeled away to allow visualization of cerebellar activation. Activation time-courses (mean percent signal change) are shown for sensorimotor regions that were unique to the V and AV modalities; time zero indicates the onset of a 6s stimulation block. Yellow: Right pSTS. Teal: Left pSTS/MTG. Magenta: Left insula.
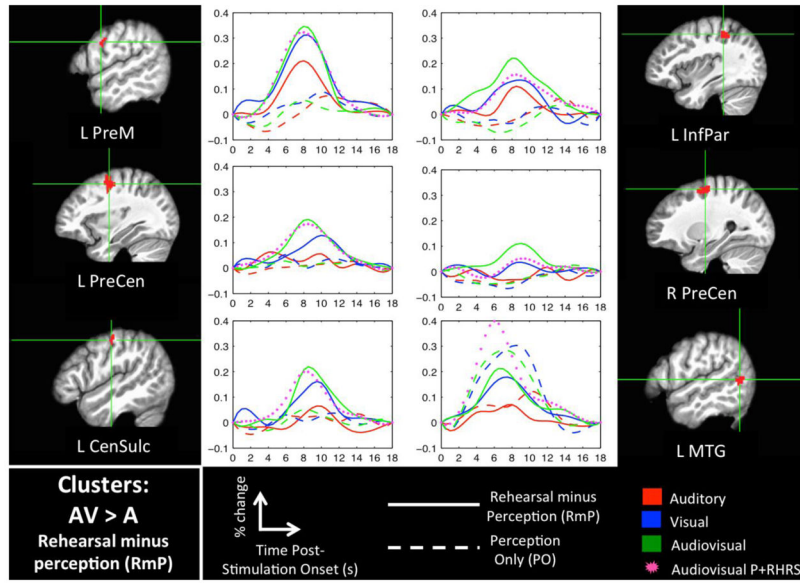
Figure 3. Interaction Analysis: AVvsA-RmP

Clusters that were significant (FWER < 0.05) in the AVvsA-RmP interaction contrast (AV-RmP > A-RmP) and also in the AV-RmP contrast alone are shown in volume space along with mean activation time-courses (percent signal change). Time zero indicates the onset of a 6s stimulation block. Also plotted is a time-course for the AV-P+RHRS condition, which is intended to show activation to covert rehearsal relative to baseline in the appropriate sensory modality.

Figure 4. Interaction Analysis: VvsA-RmP

Clusters that were significant (FWER < 0.05) in the VvsA-RmP interaction contrast (V-RmP > A-RmP) and also in the V-RmP contrast alone are shown in volume space along with mean activation time-courses (percent signal change). Time zero indicates the onset of a 6s stimulation block. Also plotted is a time-course for the V-P+RHRS condition, which is intended to show activation to covert rehearsal relative to baseline in the appropriate sensory modality.
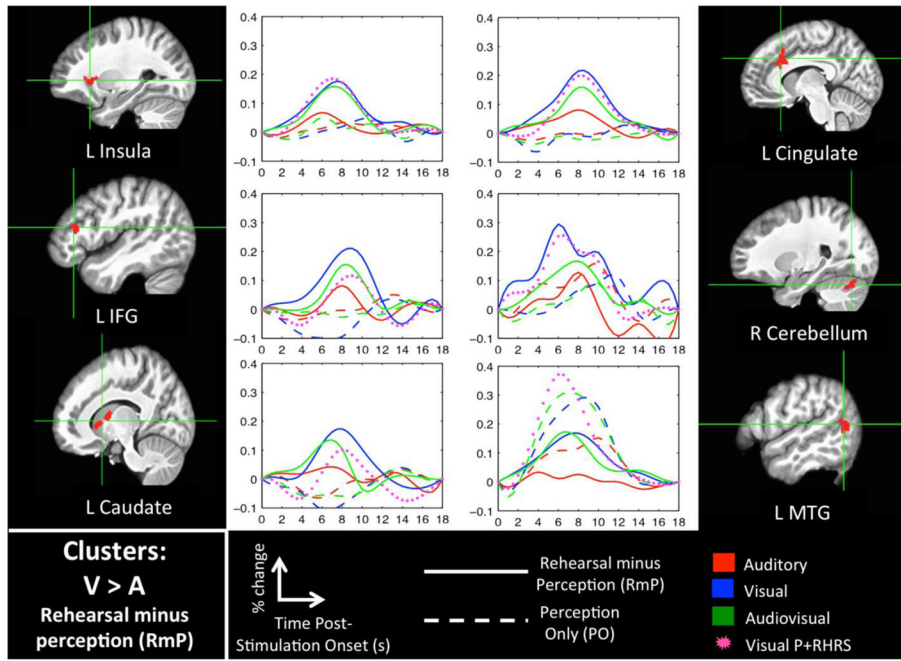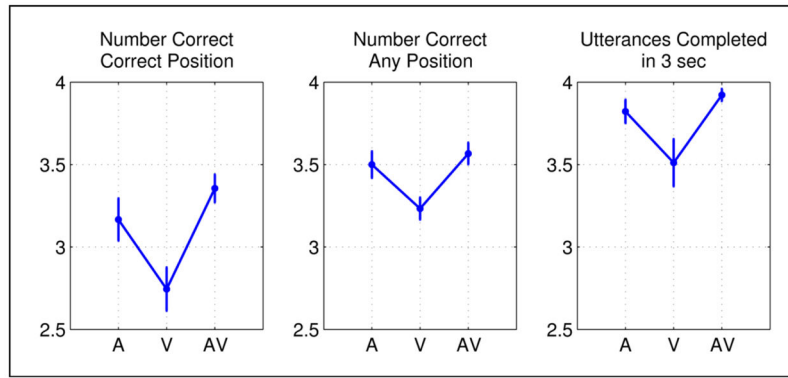
Figure 5. Results of follow-up behavioral experiment

In a behavioral experiment conducted outside the scanner, participants were asked to perceive and overtly repeat four-syllable sequences presented in three modalities – auditory (A), visual (V) and audiovisual (AV). Performance was quantified with three separate measures: number of syllables correctly repeated per trial in the correct position of the sequence, number of syllables correctly repeated per trial regardless of position, and total number of utterances per trial completed within three seconds of the onset of the response period. Means across participants are presented for each performance measure in the left, middle, and right panels, respectively. Error bars reflect ±1 SEM.

Table 1

Centers of mass (MNI) of significant clusters in Sensorimotor conjunction maps (FWER < 0.05)

	Region	Hemisphere	x	y	z	Vol (voxels)	Approximate Cytoarch. area
A SM-CONJ	SMA	L	−2.2	6.5	64.1	450	6
	Spt	L	−59.6	−42.3	22.9	364	IPC (PF)
	Cerebellum	R	29.2	−61.7	−24.6	323	Lobule VI
	PreM	L	−55.5	−3.4	49.5	320	6
	PreM	R	57.7	0.9	44.6	271	6
	IFG	L	−52.8	9.4	−1.6	251	45
V SM-CONJ	PreM	L	−54.2	−0.1	46.3	927	6
	Spt/STS	L	−56.3	−47.5	16.3	782	IPC (PF)
	SMA	L	−1.9	7.3	62.7	734	6
	PreM	R	56.9	0.8	45.3	458	6
	STS	R	54.8	−36.5	9.6	395	n/a
	Cerebellum	R	37.6	−64.3	−25.9	391	Lobule VI
	Cerebellum	L	−40.5	−65.7	−26.6	348	Lobule VIIa Crus I
	Insula	L	−35.7	22.6	3.2	313	n/a
AV SM-CONJ	Spt/STS	L	−58.8	−44.3	17.1	1271	IPC(PF)
	IFG/Insula/vPreM	L	−47.9	15.2	6.2	1028	44
	SMA	L	−2	5.8	64	690	6
	PreM	L	−53.4	−2.5	49.4	525	6
	PreM	R	57.5	0.1	44	342	6
	STS	R	47.1	−37.8	6	337	n/a
	Cerebellum	R	36.1	−65.6	−25.1	214	Lobule VI
	Putamen	L	−22	4.8	6.4	192	n/a

Table 2

Centers of mass (MNI) of significant clusters in interaction contrast maps inclusively masked with individual-modality RmP maps (each thresholded FWER < 0.05)

	Region	Hemisphere	x	y	z	Vol (voxels)	Approximate Cytoarch. area
AVvsA-RmP AND AV-RmP	PreCen Sulcus	L	−32.6	−4.6	53.8	282	6
	PreCen Sulcus	R	27.3	−9.1	59.7	251	6
	Inf Par Lobule	L	−37	−38.9	50.5	86	2
	MTG	L	−57.1	−59.9	8.8	58	n/a
	PreM	L	−58.5	1.2	33.6	55	6
	Cen Sulcus	L	−50.9	−14.2	52.2	38	1
VvsA-RmP AND V-RmP	ACC	L	−7.9	23.8	33.7	177	n/a
	Caudate Nucl.	L	−12.5	12.4	6.2	123	n/a
	Cerebellum	R	28.7	−71.9	−25.5	117	Lobule VIIa Crus I
	MTG	L	−57.5	−60.1	12.8	110	n/a
	Insula	L	−27	23.3	3.3	84	n/a
	IFG	L	−47.3	32.4	17	70	45