NEURAL BASIS OF THE CROSSMODAL CORRESPONDENCE BETWEEN AUDITORY PITCH AND VISUOSPATIAL ELEVATION

Kelly McCormick¹,³, Simon Lacey¹, Randall Stilla¹, Lynne C. Nygaard³, K. Sathian¹,²,³,⁴,⁵
¹Department of Neurology, Emory University, Atlanta, GA 30322, USA
²Department of Rehabilitation Medicine, Emory University, Atlanta, GA 30322, USA
³Department of Psychology, Emory University, Atlanta, GA 30322, USA
⁴Center for Visual and Neurocognitive Rehabilitation, Atlanta VAMC, Decatur, GA 30033, USA

Abstract

Crossmodal correspondences refer to associations between otherwise unrelated stimulus features in different sensory modalities. For example, high and low auditory pitches are associated with high and low visuospatial elevation, respectively. The neural mechanisms underlying crossmodal correspondences are currently unknown. Here, we used functional magnetic resonance imaging (fMRI) to investigate the neural basis of the pitch-elevation correspondence. Pitch-elevation congruency effects were observed bilaterally in the inferior frontal and insular cortex, the right frontal eye field and right inferior parietal cortex. Independent functional localizers failed to provide strong evidence for any of three proposed mechanisms for crossmodal correspondences: semantic mediation, magnitude estimation, and multisensory integration. Instead, pitch-elevation congruency effects overlapped with areas selective for visually presented non-word strings relative to sentences, and with regions sensitive to audiovisual asynchrony. Taken together with the prior literature, the observed congruency effects are most consistent with mediation by multisensory attention.

Keywords

multisensory; magnitude; semantic; congruency effect; attention; working memory

1. INTRODUCTION

Crossmodal correspondences are near-universally experienced associations between apparently arbitrary stimulus features in different senses (Spence, 2011). For example, large
and small visual size are consistently associated with low- and high-pitched sounds, respectively (Gallace & Spence, 2006; Evans & Treisman, 2010); and auditorily presented pseudowords, e.g., ‘takete’ and ‘maluma’, with pointed and rounded visual shapes, respectively (Köhler, 1929, 1947). A particularly well-known example of a crossmodal correspondence is that in which high and low auditory pitch are associated with high and low visuospacial elevation, respectively (e.g., Bernstein & Edelstein, 1971; Ben-Artzi & Marks, 1995; Evans & Treisman, 2010; Lacey et al., 2016; Jamal et al., 2017). Crossmodal correspondences often occur between stimulus properties that are correlated in nature. Thus, they could render information processing more efficient and facilitate integrating sensory data into unified representations (Spence, 2011). However, in some cases, crossmodal correspondences may lead to false cues: for example, although formant frequencies of animal vocalizations are inversely related to body size, many species can make atypically low sounds as a defensive strategy to exaggerate their perceived size (Fitch, 2000). The neural basis for crossmodal correspondences is unknown. Here, we focus on the audiovisual pitch-elevation correspondence and its relation to three postulated mechanisms: semantic processing, magnitude estimation, and multisensory integration.

One hypothesis for the pitch-elevation correspondence is that it may be semantically mediated because, at least in Western culture, the words ‘high’ and ‘low’ can describe both elevation and pitch (Spence, 2011; Walker et al., 2012). Some studies support this suggestion of polysemy: for instance, Shor (1975) reported Stroop interference between the auditorily presented words “high” or “low” and tones that were high- or low-pitched. In addition, these words, in relation to pitch and elevation, generated Garner interference and congruency effects for all combinations of these high/low pairs of words, tones and elevations (Ben-Artzi & Marks, 1999). Such polysemic mediation might ultimately reflect involvement of perceptual simulations evoked by the presented words, as proposed by grounded cognition accounts (e.g. Barsalou, 2008). However, other studies argue against the semantic mediation hypothesis: some non-Western cultures show the pitch-elevation correspondence despite not using spatial language to describe auditory pitch (e.g., Parkinson et al., 2012), and pre-linguistic infants are sensitive to the pitch-elevation correspondence (Walker et al., 2010; Dolscheid et al., 2014). Nonetheless, the hypothesis of semantic mediation of at least some crossmodal correspondences bears testing. If this explanation is valid, we might expect to see activity in the left hemisphere lexical-semantic network (Fedorenko et al., 2010, 2011) during processing of the pitch-elevation correspondence.

Alternatively, the pitch-elevation correspondence may arise from statistical regularities in the natural environment, given that higher-pitched sounds tend to emanate from higher spatial locations than lower-pitched sounds, which tend to occur at low spatial locations (Cabrera & Morimoto, 2007; Parise et al., 2014). This may be related to the strong tendency for vertical localizations of sounds to be strongly biased by their frequency (Cabrera & Morimoto, 2007; Roffler & Butler, 1968; Parise et al., 2014; Pratt, 1930). Thus, the crossmodal correspondence of auditory pitch and visual elevation might be related to multisensory integration based on the crossmodal co-occurrence of sounds and their visualized sources. In support of this view is our recent finding that the crossmodal pitch-elevation correspondence interacts strongly with auditory elevation (Jamal et al., 2017). This suggests that the pitch-elevation correspondence might have its roots in multisensory integration of these naturally
occurring signals. If so, neural activity underlying this crossmodal correspondence might co-localize with activity related to multisensory integration, e.g. in the superior temporal sulcus (STS) when audiovisual synchrony (Beauchamp, 2005a,b; van Atteveldt et al., 2007; Stevenson et al., 2010; Marchant et al., 2012; Noesselt et al., 2012; Erickson et al., 2014) or audiovisual identity (Sestieri et al., 2006; Erickson et al., 2014) are manipulated, or in regions such as the intraparietal sulcus (IPS) when audiovisual spatial congruency is manipulated (Sestieri et al., 2006).

Finally, certain crossmodal correspondences may have their basis in representations of magnitude (Lourenco & Longo, 2011). On this account, crossmodal correspondences could arise because the features involved, for example, size and loudness, involve polar dimensions of magnitude where one end is ‘more than’ the other (Smith & Sera, 1992). Thus, high pitch and high elevation may be associated because they are both on the same end of a polar dimension. In this case, we would expect to see activity related to the crossmodal pitch-elevation correspondence in the IPS, an area involved in processing both numerical and non-numerical (e.g., luminance) magnitude (Sathian et al., 1999; Eger et al., 2003; Walsh, 2003; Pinel et al., 2004; Piazza et al., 2004, 2007; Sokolowski et al., 2017).

The approach we took in the present study was to use functional magnetic resonance imaging (fMRI) to investigate cerebral cortical localization of congruency effects related to the crossmodal correspondence between auditory pitch and visual elevation. In order to test the relevance of the proposed mechanisms outlined above, we conducted three independent localizers in the same individuals: a semantic localizer, a magnitude estimation localizer and a multisensory synchrony localizer.

2. METHODS

2.1 Participants

Twenty participants took part in this study but two were later excluded for excessive movement in the scanner (> 1.5mm), leaving a final sample of 18 (9 male, 9 female; mean age 24 years, 9 months). All were right-handed based on the validated subset of the Edinburgh handedness inventory (Raczkowski et al., 1974) and reported normal hearing and normal, or corrected-to-normal, vision. All participants gave informed consent and were compensated for their time. All procedures were approved by the Emory University Institutional Review Board.

2.2 Procedures

2.2.1 General—Participants performed the pitch-elevation congruency scans first, and then underwent three localizer scans to test potential mechanisms underlying the pitch-elevation correspondence. After these scans, they performed a behavioral task to determine the strength of their crossmodal pitch-elevation correspondence. This fixed order was followed to avoid potential priming effects of the localizer and behavioral tasks on the pitch-elevation scans. The order of the localizer tasks was also fixed, progressing from the one perceived as most difficult to the easiest: participants did the magnitude estimation localizer first, then the temporal synchrony localizer, and finally the semantic localizer. Each localizer comprised
two runs with a fixed stimulus order; the order of runs was counterbalanced across participants. Nine out of 18 participants completed the pitch-elevation and localizer scans in a single session. The other 9 took part in a pitch-size study as well as the current pitch-elevation study; for these participants, experimental and localizer scans were performed in separate sessions (the experimental scans were first in all cases) and the intersession interval was approximately 1-2 days. All experiments were presented via Presentation software (Neurobehavioral Systems Inc., Albany CA) which allowed synchronization of scan acquisition with experiments and also recorded responses and response times (RTs). Behavioral data were analyzed in IBM SPSS v23 (IBM Corporation, Armonk NY) and effect sizes (Cohen’s d) were calculated using the online tool provided by Lenhard & Lenhard (2016); for ease of comparison between t and F statistics, the partial $\eta^2$ values provided by SPSS for effect sizes in ANOVA were transformed to Cohen’s d using the same tool.

### 2.2.2 Pitch-elevation fMRI task—
The auditory stimuli were low- or high-pitched pure tones (180Hz and 1440Hz respectively) of 200ms duration with a 20ms on/off ramp. The visual stimulus was a gray circle (RGB values 240, 240, 240) subtending approximately 1° of visual angle with its center approximately 4.2° above (high) or below (low) a central fixation cross. These stimuli were combined to form audiovisual triplets of 1000ms duration (200ms on, 200ms off), comprising three repetitions of identical stimuli that were either congruent (high pitch/high elevation or low pitch/low elevation) or incongruent (high pitch/low elevation or low pitch/high elevation: Figure 1a) with respect to the crossmodal pitch-elevation correspondence. A mirror angled over the head coil enabled participants to see the visual stimuli projected onto a screen placed in the rear magnet aperture. Auditory stimuli were presented via scanner-compatible headphones.

Because perceived loudness varies with frequency (Moore, 2012; Suzuki & Takeshima, 2004), we matched the high- and low-pitched stimuli for perceived loudness. Once positioned in the scanner and fitted with earplugs and headphones, each participant listened to the high-pitched tone at a range of amplitudes and selected the loudest tone that was still comfortable. The selected high-pitched tone was then compared to a range of low-pitched tones similarly varying in amplitude; participants selected the low-pitched tone that they perceived as matching the high-pitched tone in loudness. Participants chose high-pitched tones that ranged from approximately 95 to 102 dB SPL and low-pitched tones ranging from approximately 85 to 92 dB SPL at the headphones, before the 33 dB noise reduction by the earplugs. The high-pitched tones were, on average, 10 dB SPL greater in intensity than the low-pitched tones, but were judged equally loud by participants; while this intensity-frequency relationship is the opposite of the typical one, it presumably reflects filtering properties of the earplugs. To address the concern raised by an anonymous reviewer that scanner noise might have interfered with presentation of the auditory tones, we examined a spectrogram of the scanner noise using PRAAT v6.0.29 (Boersma & Weenink, 2008). Although the relative intensity of the scanner noise (before any attenuation by headphones and earplugs) was higher at 1440Hz (89.6dB) than at 180Hz (81.7dB), participants chose higher presentation levels for the high-pitched tones compared to the low-pitched tones.
when matching these stimuli for perceived loudness. Thus, overall, we conclude that the scanner noise did not differentially interfere with the perception of the auditory tones.

There were four runs in a slow event-related design. Each run contained 20 congruent trials (10 high pitch/high elevation and 10 low pitch/low elevation) and 20 incongruent trials (10 high pitch/low elevation and 10 low pitch/high elevation), presented in pseudorandom order with a 7s blank inter-trial interval, and 5 baseline periods of 10s duration which occurred at the start of a run, every ten trials thereafter, and at the end of a run. Total run duration was 370s. Participants performed a one-back same/different detection task by pressing a button on a hand-held response box. ‘Same’ trials, i.e. the same audiovisual combination occurring in two consecutive trials, constituted 25% of trials in each run.

2.2.3 Localizer tasks

2.2.3.1 Semantic localizer: In this task, adapted from Fedorenko et al. (2010), we contrasted complete semantically and syntactically intact sentences with strings of non-words (Figure 1b) to identify brain regions processing word- and sentence-level meaning (Fedorenko et al., 2010, 2011; Bedny et al., 2011). The complete sentences and non-word strings were 12 words long, each word or non-word being presented visually for 450ms for a total of 5.4s, followed by a 600ms inter-stimulus interval (ISI) during which participants were visually prompted to press a button. There were two runs, each consisting of 16 task blocks (8 of each type, each block containing 3 stimuli) each lasting 18s and alternating with 17 rest blocks each lasting 12s; total run duration was 492s. We expected the contrast of complete sentences > non-words to reveal regions mediating both semantic and syntactic processing (Fedorenko et al., 2010, 2011; Bedny et al., 2011), i.e., a largely left hemisphere network comprising the inferior frontal gyrus (IFG), inferior parietal cortex, and extensive sectors of the temporal lobe including the superior temporal sulcus (STS).

2.2.3.2 Multisensory integration: Among a number of possible localizers that could be used to test multisensory integration, we chose one that is sensitive to the synchrony of auditory and visual stimuli, as used in many studies of audiovisual integration (e.g., Beauchamp, 2005a,b; van Atteveldt et al., 2007; Stevenson et al., 2010; Marchant et al., 2012; Noesselt et al., 2012; Erickson et al., 2014). The auditory stimulus was an 810Hz tone of 800ms duration with a 20ms on/off ramp; this frequency was chosen to be intermediate between the 180Hz and 1440Hz tones used in the main experiment. The visual stimulus was a gray circle (RGB values 240, 240, 240) subtending approximately 1° of visual angle and presented centrally, i.e. intermediate between the two elevations used in the main experiment, for 800 ms. In synchronous trials, auditory and visual stimuli were presented simultaneously for 800ms followed by a 3200ms ISI, while asynchronous trials contained stimuli offset by an intervening blank interval of 200ms followed by a 2200ms ISI (Figure 1d). Half the asynchronous trials presented the auditory stimulus first and half the visual stimulus first. There were two runs, each consisting of 12 active blocks (6 of each type, each block containing 4 trials) each lasting 16s and alternating with 13 rest blocks each lasting 14s; total run duration was 374s. Participants had to press a button whenever an oddball stimulus (e.g., Crottaz-Herbette & Menon, 2006), either a square or a burst of white noise, occurred; two oddballs of each type occurred in each run, one in a synchronous block and
one in an asynchronous block. The contrast between synchronous and asynchronous trials was used to identify brain regions sensitive to audiovisual synchrony; we anticipated that this contrast would activate the STS and adjacent regions of the superior temporal gyrus (STG; Beauchamp, 2005a,b; van Atteveldt et al., 2007; Stevenson et al., 2010; Marchant et al., 2012; Noesselt et al., 2012; Erickson et al., 2014).

2.2.3.3 Magnitude estimation: In order to identify brain regions sensitive to magnitude, we used a modified form of the estimation task from Lourenco et al. (2012). In each trial of this task, participants were asked to estimate whether there were more black or white elements in a visual array of small oblong shapes. For a control task, we modified these arrays so that one item was a triangle and participants indicated whether the triangle was black or white, thus the response – black or white – was the same throughout (Figure 1c). There were two runs, each containing 12 active blocks (6 of each type, each block containing 4 stimuli displayed for 1s with a 3s ISI) each lasting 14s; total run duration was 374s. The contrast of magnitude estimation > control was used to identify those regions sensitive to magnitude, which we expected would be primarily in posterior parietal cortex, particularly the IPS (Walsh, 2003; Lourenco & Longo, 2011; Eger et al., 2003; Pinel et al., 2004; Piazza et al., 2004, 2007; Sathian et al., 1999; Sokolowski et al., 2017).

2.2.4 Post-scan behavioral testing—As a final step, we tested whether participants reliably demonstrated the crossmodal pitch-elevation correspondence using the implicit association test (IAT; Greenwald et al., 1998; Parise & Spence, 2012; Lacey et al., 2016). Originally devised as a test of social attitudes (Greenwald et al., 1998), the IAT has been successfully used to test the very different associations involved in crossmodal correspondences (Parise & Spence, 2012; Lacey et al., 2016). The underlying principle is the same: response times (RTs) are faster if the stimuli assigned to a particular key are congruent and slower if they are incongruent (Greenwald et al., 1998; Parise & Spence, 2012). The advantage of the IAT for testing crossmodal correspondences is that presenting each stimulus in isolation eliminates the confound of selective attention effects potentially causing slower RTs for incongruent pairings (Parise & Spence, 2012).

The auditory and visual stimuli were the same as for the imaging experiment except that stimulus duration was 1000ms and the auditory stimuli had a 100ms on/off ramp. As in the imaging experiment, each participant listened to the high-pitched tone at a range of amplitudes and selected the loudest tone that was still comfortable. This high-pitched tone was then compared to a range of low-pitched tones similarly varying in amplitude; participants selected the low tone that they perceived as matching the high tone in loudness in a procedure similar to that employed in the scan session (see above). Although the individual amplitude levels were not recorded, the auditory stimuli were presented at a comfortable listening level of approximately 73 dB SPL on average. The IAT was presented via Presentation software (Neurobehavioral Systems Inc., Albany CA) which also recorded RTs. Participants were instructed to associate pairs of stimuli with one of two response keys (the ‘left’ and ‘right’ arrow keys on a normal US ‘QWERTY’ keyboard). The pairs always consisted of one auditory and one visual stimulus and, in separate blocks of trials, were
either congruent or incongruent. A trial consisted either of an auditory presentation (a high-
or low-pitched tone) or a visual presentation (the gray circle at a high or low position) and participants were asked to respond as quickly as possible.

There were four runs, each comprising 96 trials divided into two blocks of 48 (totaling 384 trials across runs). In one block, the response key associations were congruent and in the other they were incongruent; this order was counterbalanced across runs. Each congruent/incongruent block was preceded by an instruction screen describing the response key associations to be used, and by 12 practice trials (not included in the analysis) with on-screen feedback as to accuracy (for the practice trials only). For two runs, the congruent pairs were high pitch/high position (to be associated with the left arrow key) and low pitch/low position (associated with the right arrow key) and the incongruent pairs were high pitch/low position (left arrow key) and low pitch/high position (right arrow key). On the other two runs, the left/right key associations were reversed. Half the trials were auditory (high/low pitch) and half were visual (high/low position), split equally into congruent and incongruent blocks and occurring in pseudorandom order within each block. Trials consisted of a blank 1000ms followed by either a visual or an auditory stimulus for 1000ms and were terminated either by the participant pressing a response key or automatically 3500ms after stimulus onset. RTs were measured from stimulus onset. The length of each active block thus varied between participants but was a maximum of 330s.

### 2.3 Image acquisition

MR scans were performed on a 3 Tesla Siemens Trio TIM whole body scanner (Siemens Medical Solutions, Malvern, PA), using a 12-channel head coil. T2*-weighted functional images were acquired using a single-shot, gradient-recalled, echoplanar imaging (EPI) sequence for BOLD contrast. For all functional scans, 34 axial slices of 3.1mm thickness were acquired using the following parameters: repetition time (TR) 2000ms, echo time (TE) 30ms, field of view (FOV) 200mm, flip angle (FA) 90°, in-plane resolution 3.125×3.125mm, and in-plane matrix 64×64. High-resolution 3D anatomic images were acquired using an MPRAGE sequence (TR 2300ms, TE 3.9ms, inversion time 1100ms, FA 8°) comprising 176 sagittal slices of 1mm thickness (FOV 256mm, in-plane resolution 1×1mm, in-plane matrix 256x256). Once magnetic stabilization was achieved in each run, the scanner triggered the computer running Presentation software so that the sequence of experimental trials was synchronized with scan acquisition.

### 2.4 Image processing and analysis

Image processing and analysis was performed using BrainVoyager QX v2.8.4 (Brain Innovation, Maastricht, Netherlands). In individual analyses, each participant’s functional runs were real-time motion-corrected utilizing Siemens 3D-PACE (prospective acquisition motion correction). Functional images were preprocessed utilizing cubic spline interpolation for slice scan time correction, trilinear-sinc interpolation for intra-session alignment of functional volumes, and high-pass temporal filtering to 2 cycles per run to remove slow drifts in the data without affecting task-related effects. Anatomic 3D images were processed, co-registered with the functional data, and transformed into Talairach space (Talairach & Tournoux, 1988). Talairach-normalized anatomic data sets from multiple scan sessions (2-3
per participant) were averaged for each individual, to minimize noise and maximize spatial resolution.

For group analyses, the transformed data were spatially smoothed with an isotropic Gaussian kernel (full-width half-maximum 4mm). The 4mm filter is within the 3-6mm range recommended to reduce the possibility of blurring together activations that are in fact anatomically and/or functionally distinct (White et al., 2001). Runs were percent signal change normalized (i.e., the mean signal value for each voxel’s time course was transformed to a value of 100, so that the individual values fluctuated around that mean as percent signal deviations).

For group activation display and statistical analyses, we created a group average brain. We first selected a representative (target) Talairach-normalized brain from the 18-participant group. We then individually aligned the 17 remaining participants’ Talairach-normalized brains to this target (co-registration to match the gyral and sulcal pattern, followed by sinc interpolation). These 17 aligned brains were then averaged. This 17-subject average brain was then averaged with the target brain, creating a single Talairach template, with 1mm isotropic resolution, which was used to display the activations for the 18-subject group. This 18-subject Talairach template was then manually segmented, creating a group average cortical ‘mask’ file with 3mm spatial resolution, equivalent to the spatial resolution of the functional data files. This group average brain was displayed using the real-time volume rendering option in BrainVoyager QX.

Statistical analyses of group data used general linear models (GLMs) treating participant as a random factor (so that the degrees of freedom equal n-1, i.e. 17), followed by pairwise contrasts. This analysis allows generalization to untested individuals. Correction for multiple comparisons within a cortical mask (corrected p < .05) was achieved by imposing a threshold for the volume of clusters comprising contiguous voxels that passed a voxel-wise threshold of p < 0.001, using a 3D extension (implemented in BrainVoyager QX) of the 2D Monte Carlo simulation procedure described by Forman et al. (1995). This stringent voxel-wise threshold is recommended to avoid potential problems of false positives and also permits more accurate spatial localization of activations than when more liberal thresholds are used (Woo et al., 2014; Eklund et al., 2016). Activations were localized with respect to 3D cortical anatomy with the help of an MRI atlas (Duvernoy, 1999).

In reporting activations, we do not provide ‘hotspot’ coordinates (i.e. the voxel with the largest t-value) because the statistical significance of specific voxels is not tested against other voxels within the activation (Woo et al., 2014). Instead, we provide the ‘center of gravity’ (CoG) coordinates since these orient the reader to the anatomical location but are statistically neutral. Where an activation spans several anatomical locations, we provide an extended description (see Tables 1-4). Likewise, in order to compare the present results to previous studies, we have visually inspected activations reported in those papers rather than compute the Euclidean distance between coordinates since this would have involved relying on ‘hotspot’ coordinates in most cases.
3. RESULTS

3.1 Behavioral

3.1.1 In-scanner tasks—In the semantic localizer, participants correctly responded (mean ± sem) to the visual cue at the end of each sentence or non-word list in 98.5±0.8% of trials on average. For the multisensory integration localizer, participants correctly identified 88.9±4.2% of the oddball trials on average. For the magnitude localizer, there was no significant difference in accuracy between the magnitude estimation (94.7±1.8%) and control (96.9±1.2%) tasks (t_{17} = −1.02, p = .3, d = .34) although RTs were significantly faster for the control task (981±60ms) compared to the magnitude task (1109±64ms; t_{17} = 4.53, p < .001, d = .48).

Pitch-elevation task: Mean (±sem) accuracy was not significantly different between trials requiring a “same” response (93.4±2.3%) and those requiring a “different” response (95.0±1.9%: t_{17} = −.56, p = .6, d = .19), nor between congruent (94.5±1.6%) and incongruent (94.8±1.5%) trials (t_{17} = −.45, p = .7, d = .05). Although there were more “different” trials than “same” trials (75% vs. 25%), potentially leading to response bias, the false alarm rates for both congruent and incongruent trials were quite low (5.5% vs. 5.2% respectively) and hence we did not attempt to correct for response bias by computing d. RT analyses excluded trials for which there was no response (1.3% of all trials), incorrect responses (5.3% of responses), and further excluded trials for which the RT was ± 2.5 standard deviations from the individual participant mean (3.1% of correct response trials). RTs were not significantly different between congruent (1197±75ms) and incongruent (1199±75ms) trials (t_{17} = −.19, p = .8, d < .01).

3.1.2 Post-scan pitch-elevation IAT—Repeated-measures ANOVA (RM-ANOVA) with factors of response key association (congruent, incongruent) and stimulus modality (auditory, visual) showed that accuracy was higher for congruent (96.1±0.4%) than incongruent associations (91.5±1.5%: F_{1,17} = 10.2, p = .005, d = 1.55) but there was no main effect of modality (F_{1,17} = 1.1, p = .7, d = .18) and no interaction between these factors (F_{1,17} = 1.7, p = .2, d = .64). RT analyses excluded trials for which there was no response, or which failed to log (.62% of all trials), incorrect responses (6.2% of responses), and further excluded trials for which the RT was ± 2.5 standard deviations from the individual participant mean (2.9% of correct response trials). RM-ANOVA showed that RTs were faster when the response key associations were congruent (553±23ms) compared to incongruent (695±37ms: F_{1,17} = 28.5, p < .001, d = 2.59). RTs were also faster for visual (537±26ms) compared to auditory (711±32ms) stimuli (F_{1,17} = 105.8, p < .001, d = 5.00). The response key association × stimulus modality interaction was not significant (F_{1,17} = 3.9, p = .07, d = .95). Only one participant failed to show a congruency effect, i.e. faster RTs for congruent trials, and only for auditory stimuli.

3.2 Imaging

3.2.1 Localizer tasks

3.2.1.1 Semantic: The contrast of complete sentences > non-words within the cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 8 voxels)
revealed, as expected, a predominantly left hemisphere network that included the pars triangularis of the IFG and a large activation along the STS that extended into the superior temporal gyrus (STG), the latter activation being bilateral (Table 1a; Supplementary Figure 1). Thus, the semantic localizer essentially replicated the previously reported language network (Fedorenko et al., 2010, 2011).

We also examined the reverse contrast of non-words > complete sentences which, to our knowledge has not been reported previously but which may reflect phonological processing (Fedorenko et al., 2010; and see Discussion). This contrast, within the cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 8 voxels), revealed bilateral activations at several foci in the cingulate gyrus and sulcus, in medial superior frontal gyrus (SFG), and in the middle frontal gyrus (MFG) extending to the inferior frontal sulcus (IFS). There were additional right hemisphere activations in MFG and SFG foci distinct from those already mentioned, as well as in the angular gyrus (AG) and a large region extending from the inferior frontal gyrus (IFG) through the IFS to the MFG. Left hemisphere activations were found in the IFG extending to the lateral orbital gyrus, the parieto-occipital fissure (POF) extending to the superior parietal gyrus (SPG), the supramarginal gyrus (SMG) extending to the AG, and a large region extending from the posterior through the mid-insula into the pars opercularis of the IFG (Table 1b; Supplementary Figure 1).

3.2.1.2 Multisensory synchrony: The contrast of synchronous > asynchronous within the cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 8 voxels) revealed activations in the right superior occipital gyrus (SOG) and left POF extending to the posterior cingulate gyrus (Table 2a; Supplementary Figure 2). The reverse contrast, asynchronous > synchronous, showed two right hemisphere locations, one extending from the AG to the anterior IPS and one extending from the MFG into the IFG and IFS. There was also a single activation in the left IPS (Table 2b; Supplementary Figure 2). Note that some previous studies reported greater activation for synchronous compared to asynchronous audiovisual stimuli, whereas others found the reverse to be the case; in some cases this was true of foci in proximity to each other (e.g. Stevenson et al., 2010). Regions sensitive to synchrony were reported, in earlier studies, most consistently in the STS and/or STG – these regions did not emerge in the present study from the comparison of synchronous to asynchronous audiovisual stimuli, perhaps reflecting differences in stimuli and tasks (see Discussion). However, the right SOG focus that showed greater activation in the present study for synchronous relative to asynchronous stimuli was near one of the loci showing a similar preference in the study of Stevenson et al. (2010), and the right frontal region that was more active on our localizer for asynchronous than synchronous stimuli was close to a region identified on the meta-analysis of Erickson et al. (2014) as exhibiting a preference for audiovisual stimuli characterized by either content incongruency or asynchrony.

3.2.1.3 Magnitude estimation: The contrast of magnitude estimation > control within the cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 7 voxels) revealed exclusively right hemisphere activity including several locations in and
around the IPS (Table 2; Supplementary Figure 3). These loci are consistent with activations reported in previous studies of magnitude processing (Eger et al., 2003; Pinel et al., 2004; Piazza et al., 2004, 2007) and a recent meta-analysis (Sokolowski et al., 2017). An additional focus in the right middle occipital gyrus (MOG) is close to regions that were previously implicated in subitizing (Sathian et al., 1999) or that showed adaptation to magnitude (Piazza et al., 2007).

3.2.2 Pitch-elevation task—To test for regions processing the crossmodal pitch-elevation correspondence, we tested for voxels showing a greater response for congruent than incongruent trials (C > I), i.e., a congruency effect. At a voxel-wise threshold of p < .001, there were no activations that survived correction for multiple comparisons. The absence of a neural congruency effect for the C > I contrast mirrors the lack of a behavioral congruency effect when comparing all congruent trials to all incongruent trials (see above). However, cortical activity during any given trial might have been confounded by the demands of the one-back same/different task that participants performed in the scanner, which required that participants maintain the previous trial in working memory for comparison with the current trial. When consecutive trials differed in their congruency type, either congruent preceded by incongruent (IC) or vice versa (CI), the requirement to compare congruent and incongruent trials might have contaminated the activity profile for the current trial type, thus obscuring any congruency effect.

We therefore tested for a congruency effect when consecutive trials were of the same type, either both congruent (CC) or both incongruent (II). The contrast CC > II within the cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 7 voxels) revealed bilateral activation in the anterior insula, a left hemisphere region extending from the pars opercularis of the IFG to the mid-insula, and right hemisphere activations in the frontal eye field (FEF), mid-insula, anterior and frontal opercular regions of the IFG, and the AG extending to the mid-IPS (Table 4a; Figure 2). Thus, the CC > II contrast showed a congruency effect when the preceding and current trials were not in conflict. This does not imply the absence of any modulation of a given trial by the immediately preceding trial – processing of the current trial could be facilitated by a matching preceding trial – but the congruency effect was at least not obscured in this case. Note that the CC > II neural congruency effect occurred in the absence of a significant behavioral congruency effect in these trials (accuracy: CC trials 95.9±1.3%, II trials 96.5±1.1%, t_{17} = -.9, p = .4, d = .10; RTs: CC trials 1178±78ms, II trials 1185±74ms, t_{17} = -.3, p = .8, d = .02).

In contrast, when we tested for a congruency effect only in those trials where the preceding trial was of a different type (IC > CI), a blank map was obtained, supporting the idea that when consecutive trials were of different types, any congruency effect was likely eliminated by the demands of the 1-back task.

While the CC > II contrast unambiguously reveals regions that are sensitive to the pitch-elevation correspondence, it does not, as noted above, rule out all effects of the previous trial. To probe this further, we carried out the interaction contrast (CC > II) > (IC > CI) to test for the interaction of congruency effects between trials. We reasoned that any voxels from the CC > II contrast that also appeared in the interaction contrast might reflect
modulatory effects from the preceding trial. On the other hand, those voxels from the CC > II contrast that were absent from the interaction effect could be considered essentially unmodulated by the preceding trial and thus to represent the “purest” congruency effect possible with our experimental design.

The interaction contrast revealed bilateral AG activation, extending to the mid-IPS on the right, together with activations in the right anterior IFG and MFG, the left SMG and left orbital gyrus (Table 4b; Figure 2). Of these, the right anterior IFG and AG/mid-IPS activations overlapped with those from the CC > II contrast. The sensitivity of these regions to both the CC > II contrast and the interaction effect makes it difficult to specify the process(es) responsible for their activation in the present study. The remaining regions from the CC > II contrast that were insensitive to the interaction effect can be considered more strongly as loci of the pitch-elevation congruency effect. These regions included the right FEF and bilateral opercular IFG and insular foci. Conversely, parts of the left AG, SMG, and orbital gyrus, along with a focus in the right MFG, were sensitive to the interaction but not the CC > II congruency effect, i.e. these were loci involved in inter-trial modulation but not in the primary congruency effect.

The multisensory integration localizer did not show the activity in the STS that was expected on the basis of previous studies (Beauchamp, 2005a,b; van Atteveldt et al., 2007; Stevenson et al., 2010; Marchant et al., 2012; Noesselt et al., 2012; Erickson et al., 2014). Therefore, in order to test the multisensory integration hypothesis in the STS, we conducted a region of interest (ROI) analysis. We created bilateral STS ROIs consisting of 15mm cubes centered on coordinates from Werner & Noppeney (2010; MNI coordinates were transformed into Talairach space using the online tool provided by Lacadie et al., 2008). These loci were chosen because they were sensitive to audiovisual integration of non-speech stimuli (given that our stimuli were non-speech, albeit of a different type than used in this study), and were also behaviorally relevant in that activation profiles predicted task performance (Werner & Noppeney, 2010). However, the CC > II contrast within a mask of these ROIs (396 voxels) did not result in any activations.

Finally, we also tested whether activation magnitudes from the CC > II contrast correlated, across participants, with the magnitude of individual congruency effects derived from in-scanner RTs and from RTs on the post-scan IAT. The congruency effects used the formula \((RT_i - RT_c) / (RT_i + RT_c)\), where \(RT_i\) and \(RT_c\) represent RT for incongruent and congruent trials, respectively, on the post-scan IAT, and for II and CC trials, respectively, for in-scanner RTs. To avoid circularity, we conducted these correlation tests independently of the activations and in a similarly stringent manner, by setting a strict voxel-wise threshold of \(p < 0.001\) within the cortical mask before applying cluster correction (corrected \(p < 0.05\)). There were no foci within the cortical mask that showed significant correlation with either congruency effect.

### 3.2.3 Congruency effect overlap with localizers

Our approach to distinguishing between competing potential mechanisms was to look for overlap between areas showing pitch-elevation congruency effects and areas revealed by the semantic, magnitude, and multisensory localizers. Note that we compared pitch-elevation and localizer maps when
both had a voxel-wise threshold of p < .001. Overlaps at this strict threshold, avoiding potential false positives and allowing more accurate spatial localization than at more liberal thresholds (Woo et al., 2014; Eklund et al., 2016), would provide support for candidate mechanisms. However, note that absence of overlaps would not allow such mechanisms to be definitively ruled out. It is worth reiterating that the right anterior IFG and AG/mid-IPS congruency regions, where they occur in this analysis, were also sensitive to the interaction effect and may not be “pure” congruency regions.

None of the regions showing a congruency effect overlapped, or were contiguous with (i.e. shared a common edge or vertex), any area revealed by the main localizer contrasts, i.e. those intended to reflect potential mechanisms (Figure 3). Rather, regions showing pitch-elevation congruency effects showed overlaps and contiguities with the control conditions from the semantic and multisensory localizers, i.e. when the primary contrasts for these conditions were reversed (see notes to Table 4a; Figure 3). This was unexpected. The contrast of non-words > complete sentences from the semantic localizer revealed overlap with congruency effects in the pars opercularis of the left IFG/mid-insula. In addition, the right AG/mid-IPS cluster from the congruency effect showed overlap of its AG part, and contiguity of its IPS part, with a region exhibiting preference for non-words over sentences. Also, the contrast of asynchronous > synchronous from the multisensory localizer showed contiguity with the right AG/mid-IPS congruency region.

4. DISCUSSION

The present study investigated the neural basis of congruency effects related to the crossmodal correspondence of auditory pitch and visuo-spatial elevation. We used independent localizers to examine the relationship between these neural congruency effects and three potential mechanisms: semantic mediation, magnitude estimation and multisensory integration. While a significant congruency effect was not found for behavioral data obtained during scanning, the post-scan IAT in the same individuals did demonstrate a behavioral congruency effect. The absence of a behavioral congruency effect during scanning may reflect the specific task employed, which required a one-back same-different discrimination on the audiovisual combination of stimuli used. This idea is reinforced by our imaging data: we failed to find a simple neural congruency effect by contrasting congruent with incongruent trials. Nonetheless, we did find a neural congruency effect when contrasting the subsets of congruent and incongruent trials preceded by trials of the same congruency type, with greater activation for the congruent compared to incongruent trials. Also, there was a significant interaction between the congruency type on a given trial and that on the immediately preceding trial. The right FEF and bilateral regions occupying the anterior/mid-insula and frontal operculum exhibited congruency effects uncontaminated by the interaction effect, whereas regions in the left inferior parietal cortex, left orbital gyrus and right MFG demonstrated the interaction but not the congruency effect. Foci in the inferior frontal and parietal cortex of the right hemisphere were sensitive to both the congruency and interaction effects. To our knowledge, this is the first report of the neural correlates of the pitch-elevation crossmodal correspondence.
4.1 Semantic mediation

Although the semantic localizer successfully reproduced the semantic network identified by Fedorenko et al. (2010, 2011), there was no overlap with regions showing the congruency effect for the pitch-elevation correspondence. Prima facie, this would argue against semantic mediation of the pitch-elevation correspondence, at least in terms of sentence- and word-level meaning processed by the kind of semantic network revealed by the localizer (although, as stated in the Results, the absence of overlaps does not allow us to definitively rule out a semantic explanation). It is possible that the crossmodal pitch-elevation correspondence is related to more implicit, object-level semantic associations like those between the characteristic sounds and typical images of inanimate objects or animals. Indeed, a left inferior frontal region was more strongly activated by audiovisual incongruency of familiar objects compared to audiovisual stimuli representing unfamiliar objects (see Figure 3 of Hein et al., 2007). This focus appears to be close to a cluster exhibiting a congruency effect for the pitch-elevation correspondence in the present study: the pars opercularis of the left IFG extending into the mid-insula. Note that the pars opercularis is part of Broca’s area and lies within a region sensitive to a verb > noun contrast (Tyler et al. (2004), which might be taken as suggestive of semantic mediation. These potential relationships to semantic processing emerging from comparisons with prior studies may be subject to the pitfalls of reverse inference (Poldrack, 2006, 2011), although Hutzler (2014) argued that careful evaluation of the tasks utilized may allow reverse inference to be valid. We stress that where we draw reverse inferences, these are to be treated as hypotheses to be tested in further experimental work and not as firm conclusions about underlying brain processes. Thus, the possibility of semantic mediation of the pitch-elevation correspondence is not strongly supported by the present study but may merit further investigation. Other lines of evidence argue against semantic mediation: Cross-cultural studies show that different cultures experience the same pitch-elevation correspondence independently of the linguistic terms used to describe it (Parkinson et al., 2012). The developmental literature indicates that pre-verbal infants are sensitive to the pitch-elevation correspondence (Walker et al., 2010; Dolscheid et al., 2014). Furthermore, our recent finding that the pitch-elevation correspondence interacts strongly with auditory elevation (Jamal et al., 2017) suggests its origin in the statistics of the natural world, in which high- and low-pitched sounds tend to emanate from high and low spatial locations, respectively (Cabrera & Morimoto, 2007; Parise et al., 2014). This, in turn, would point to multisensory integration of such statistically associated stimuli as a possible mechanism underlying the pitch-elevation correspondence. This possibility is addressed later.

4.2 Relationship to phonological processing

Unexpectedly, some of the clusters demonstrating the pitch-elevation congruency effect overlapped, or were contiguous, with areas sensitive to the control task used in the semantic localizer, i.e., the reverse contrast of non-words > sentences. To our knowledge, this contrast has not been reported previously, but since reading non-words depends largely on phonological processing, whereas reading sentences also involves syntactic and semantic processes, the contrast of non-words > sentences could reflect phonological processing (Fedorenko et al., 2010). Another possibility is that some or all of the activations on this contrast might be related to more effortful reading of non-words than complete sentences.
(see Price et al., 1996); in support of this idea, many of the activations on this contrast (principally those in the IFG, IFS and MFG bilaterally) were in regions identified as part of the domain-general frontoparietal system (Duncan, 2013; Fedorenko et al., 2013).

One of the clusters showing the pitch-elevation congruency effect was the activation in the pars opercularis of the left IFG extending into the mid-insula; the other was in the right AG/mid-IPS. Note that the latter cluster was also sensitive to the interaction effect between the congruency type of successive trials. As noted above, stronger responses during reading of non-words compared to sentences could reflect phonological processing (Fedorenko et al., 2010). The finding (Binder et al., 2005) of greater activity for non-words relative to concrete words in the left IFG near the activation found in the present study is consistent with a phonological basis for activation at this locus. Although it is unclear how exactly phonological processes might relate to the pitch-elevation crossmodal correspondence, phonology does involve crossmodal re-coding of visual into auditory representations (Goswami 1993, 2008); thus, it is possible that similar neural processes might underlie the audiovisual correspondence between pitch and elevation. Arguing against a relationship between phonology and crossmodal correspondence in the case of the right parietal site is that activations related to phonological processing are predominantly left hemispheric (Glezer et al., 2016; Zhang et al., 2017). Thus, given that the potential relationship to phonology was not predicted, and further since the right parietal cluster active on the congruency contrast was also sensitive to the interaction effect between the congruency type of a particular trial and its immediate predecessor, this relationship requires further experimental evaluation. The alternative possibility, also mentioned above, is recruitment of domain-general frontoparietal cortical areas due to more effortful reading of non-words than sentences (Duncan, 2013; Fedorenko et al., 2013). This fits with potential involvement of attentional control processes in the congruency-related activation in the pars opercularis of the left IFG (comparing our Figure 2 to Figure 2a in Duncan, 2013). Duncan (2013) argued that a key function of the domain-general or multiple-demand system is to represent “attentional episodes” that are critical in fractionating multi-step cognitive processes into manageable chunks. Further arguments implicating attention in the observed congruency effects are advanced in a subsequent section.

4.3 Involvement of magnitude estimation

Since the magnitude localizer replicated areas in and around the right IPS found in many previous studies (Eger et al., 2003; Pinel et al., 2004; Piazza et al., 2004, 2007; Sathian et al., 1999; Sokolowski et al., 2017), we can be quite confident in the localizer results. Congruency effects for the pitch-elevation crossmodal correspondence in the right mid-IPS were 17 mm (Euclidean CoG distance) away from IPS activation on the magnitude localizer. This may indicate, albeit not very strongly, some relatedness, if not commonality, of processing between magnitude estimation and the pitch-elevation crossmodal correspondence. The IPS is also involved in processing relative pitch, as inferred from a task requiring recognition of transposed melodies (Foster & Zatorre, 2010); this may be relevant to the pitch-elevation correspondence since this pitch-related IPS focus is close to the right mid-IPS congruency region. In addition, magnitude can be defined non-numerically, for instance brighter/darker or louder/quieter. Such sensory-based scales may be more relevant...
to the pitch-elevation correspondence than the numerosity-based task we used, although there is considerable overlap between the regions involved in number and luminance comparisons (Pinel et al., 2004). It should also be noted that one complication with the magnitude account of crossmodal correspondences is that understanding polar dimensions, and cross-dimension mappings, may depend on language acquisition and could thus be semantically mediated (Smith & Sera, 1992). In adults, perception and language converge on a coherent organization in which the perceptual poles of a particular dimension and the meanings of its associated ‘more/less’ dimension words are unified (Smith & Sera, 1992). However, as noted above, we did not find strong evidence for semantic mediation of the pitch-elevation crossmodal correspondence. Furthermore, the relationship between auditory pitch and magnitudes in other dimensions is complex; high pitch is associated with increasing height, loudness, and lightness but also with decreasing mass, quantity and size (Eitan & Timmers, 2010). Future work should compare the neural relationships between these different magnitude dimensions and different crossmodal correspondences involving auditory pitch.

### 4.4 Relationship to multisensory integration

As pointed out in the Results section, the multisensory localizer did not reveal activity in STS or nearby parts of superior temporal cortex, which was the region most consistently activated in previous studies that manipulated the synchrony of audiovisual stimuli. The differences in activations probably reflect varying stimuli and tasks: None of these earlier studies used exactly the same stimuli as used in the present study. Whereas we used a tone and a circle, most prior studies employed audiovisual speech (van Atteveldt et al., 2007; Stevenson et al., 2010; Noesselt et al., 2012; Erickson et al., 2014); one study did use tones but these were paired with checkerboards (Marchant et al., 2012). Our localizer was designed using stimuli very similar to those used in the pitch-elevation congruency analysis, with the goal of revealing multisensory processing most relevant to the pitch-elevation crossmodal correspondence. Further, we specifically avoided speech stimuli in order to steer clear of potential semantic effects, given that one of the hypotheses being tested was semantic mediation of the pitch-elevation crossmodal correspondence. Finally, there is more perceptual tolerance of asynchrony (i.e. offset stimuli are more likely to be perceived as synchronous) when the visual stimulus precedes the auditory stimulus (V-A) than vice versa (A-V: Bhat et al., 2015). In the asynchronous condition, we had equal numbers of V-A and A-V trials; therefore, to the extent that asynchronous V-A trials were actually perceived as synchronous, the effectiveness of the synchronous > asynchronous contrast may have been compromised.

There was no overlap between regions showing a congruency effect for the pitch-elevation crossmodal correspondence and those responding more strongly to synchronous, compared to asynchronous, audiovisual stimuli; nor did we find congruency effects in an ROI analysis of bilateral STS foci derived from a prior study of non-speech stimuli (Werner & Noppeney, 2010). Rather unexpectedly, the AG part of the right parietal cluster sensitive to both congruency and the interaction effect abutted, and its mid-IPS portion was close to, areas sensitive to audiovisual asynchrony. Although it is difficult to relate asynchrony to the congruency of the pitch-elevation correspondence, one possibility is that the oddball
detection task used in the multisensory localizer may have been more demanding when the auditory and visual stimuli were asynchronous, compared to when they were synchronous. This explanation is in keeping with the suggestion made above that this parietal cluster may have been recruited by effort-related processes involved in reading non-words, relative to sentences.

Although the pitch of sounds delivered from a neutral location can elicit congruency effects with visuospatial elevation, auditory spatial elevation is well known to interact with auditory pitch (Cabrera & Morimoto, 2007; Parise et al., 2014; Pratt, 1930; Roffler & Butler, 1968), and we showed recently that auditory spatial elevation significantly modulates the crossmodal correspondence between auditory pitch and visual elevation (Jamal et al., 2017). Hence, spatial rather than temporal congruency may be more relevant to the pitch-elevation crossmodal correspondence (see Spence, 2013, for the importance of spatial congruency in multisensory integration). In fact, a region of the right IPS and adjacent inferior parietal lobule, close to our right AG/mid-IPS congruency cluster, was noted to be more active during localization compared to recognition of multisensory objects, but this region was not sensitive to spatial congruency (Sestieri et al., 2006). Therefore, even this consideration of a previous spatial congruency manipulation does not provide support for the multisensory integration hypothesis of the pitch-elevation crossmodal correspondence. However, subject to caveats about reverse inference, congruency effects that we found in the right FEF are of at least some interest in a multisensory context since this region shows early responses to both visual and auditory stimuli (Kirchner et al., 2015).

4.5 Role of multisensory attention

Support for a multisensory attentional account of the pitch-elevation crossmodal correspondence is provided by our finding of a congruency effect in the anterior insula bilaterally, when other studies of this region’s activity are considered. This area has previously been shown to reflect task difficulty in audiovisual synchrony judgments of a flash and a tone, but also in other unisensory tasks (Lamichhane et al., 2016). Additionally, it has been proposed that the right anterior insula is involved in integrating audiovisual inputs during attention to multisensory stimuli (Chen et al., 2015). This region also responds preferentially to novel compared to familiar stimuli across multiple sensory modalities – auditory, visual and tactile (Downar et al., 2002), and is active during attention to audiovisual speech stimuli that are incongruent between modalities (Fernandez et al., 2015). Taken together, the task-related “reverse inference” (Hutzler, 2014) most parsimoniously emerging from consideration of these prior studies in relation to the present study is that engagement of the anterior insula may reflect processes related to multisensory attention; this possibility remains to be tested in further work. Anterior insular activity also fits with this region being part of the multiple demand network proposed by Duncan (2013).

Additional support for involvement of multisensory attention as a mediator of the effects observed here comes from the finding that the dorsal attention network, including a right FEF focus not very distant from our focus in this area, was active during a semantic congruency task; the ventral attention network, including a focus in the right AG near the AG part of our AG/mid-IPS cluster, was also recruited but only in the presence of additional top-down demands on attention (Mastroberardino et al., 2015). Our task, by its design
involving a one-back same-different discrimination of multisensory stimuli, may have strongly engaged top-down attentional processes relevant for multisensory perception, as well as bottom-up processes triggered by salience of the congruent multisensory stimuli (Quak et al., 2015). Further backing for an attentional account of the congruency effects is provided by considerations outlined earlier, i.e. overlap of the congruency effects with selectivity for non-word over sentence reading in the left frontal operculum/mid-insula and right inferior parietal cortex, and with selectivity for asynchrony over synchrony in the right inferior parietal cortex.

4.6 Involvement of working memory

Our one-back task design also probably depended on working memory. The neural congruency effect we observed for the pitch-elevation correspondence fits with the observation of Brunetti et al. (2017) that target detection in a working memory n-back task is faster for congruent compared to incongruent presentations, although a behavioral congruency effect was absent in our in-scanner data (but was present on the IAT outside the scanner). The frontal and inferior parietal regions sensitive in the present study to the interaction effect between successive trials, of which foci in right inferior frontal and parietal cortex also exhibited the pitch-elevation congruency effect, were among the regions implicated in audiovisual working memory for semantically congruent stimuli (Xie et al., 2017), with the strongest effects in this working memory study being noted in inferior parietal cortex. Although localization in the study of Xie et al. (2017) was based on source reconstruction of scalp electroencephalographic data, which tends to limit the extent to which the data can be compared with the current fMRI data, this task-based reverse inference is consistent with the idea that multisensory working memory processes generated the interaction effects in our study, and is worthy of further investigation. Moreover, the co-localization of the pitch-elevation congruency effect with the interaction effects in right inferior parietal cortex lends further weight to the proposed role of multisensory attention outlined in the preceding section, given the close interplay between multisensory processing, attention, working memory and long-term memory in the context of a predictive coding (Bayesian) framework (Quak et al., 2015).

4.7 Conclusions

This study is the first to examine the neural basis of the crossmodal correspondence of pitch and elevation. Among the a priori mechanisms we investigated, evidence for mediation of the pitch-elevation crossmodal correspondence by neural processes involved in semantics or magnitude estimation was, at best, weak, and evidence for multisensory integration was lacking. Although none of the three hypotheses that we tested can be definitively ruled out at present, we did find suggestive evidence for alternative explanations that can form the basis of future experiments. One of these is a potential relationship between crossmodal processes involved in the pitch-elevation correspondence and in phonology. The strongest explanation for the activations we observed is that our task probably engaged multisensory mechanisms involved in attention and working memory. It remains to be seen whether a different experimental design that does not rely on congruency effects would reveal lower-level neural mechanisms for the pitch-elevation correspondence. Thus, an important matter to be addressed in subsequent studies of crossmodal correspondences is to disentangle higher-
level cognitive processes from lower-level sensory factors mediating the correspondences,
recognizing that the relevant mechanisms might differ depending on the specific
 correspondence being tested.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**Acknowledgments**

This work was supported by grants to KS and LN from the National Eye Institute at the NIH (R01EY025978) and
the Emory University Research Council, and to KMcc from the Emory University Facility for Education and
Research in Neuroscience and the Laney Graduate School. Support to KS from the Veterans Administration is also
acknowledged. We thank Lawrence Barsalou, Justin Bonny, Daniel Dilks, Evelina Fedorenko, Harold Gouzoules,
Stella Lourenco and Kate Pirog Revill for their advice and assistance.

**Abbreviations**

<table>
<thead>
<tr>
<th>Directional</th>
<th>Anatomical</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>AG</td>
</tr>
<tr>
<td>med</td>
<td>AOS</td>
</tr>
<tr>
<td>p</td>
<td>calcS</td>
</tr>
<tr>
<td>v</td>
<td>CS</td>
</tr>
<tr>
<td>anterior</td>
<td>angular gyrus</td>
</tr>
<tr>
<td>medial</td>
<td>anterior occipital sulcus</td>
</tr>
<tr>
<td>posterior</td>
<td>calcarine sulcus</td>
</tr>
<tr>
<td>ventral</td>
<td>central sulcus</td>
</tr>
<tr>
<td>anterior</td>
<td>cingulate gyrus</td>
</tr>
<tr>
<td>posterior</td>
<td>cingulate sulcus</td>
</tr>
<tr>
<td>posterior</td>
<td>collateral sulcus</td>
</tr>
<tr>
<td>posterior</td>
<td>fusiform gyrus</td>
</tr>
<tr>
<td>frontal</td>
<td>frontal operculum</td>
</tr>
<tr>
<td>operculum</td>
<td></td>
</tr>
<tr>
<td>insula</td>
<td></td>
</tr>
<tr>
<td>inferior</td>
<td>IOG</td>
</tr>
<tr>
<td>occipital</td>
<td></td>
</tr>
<tr>
<td>gyrus</td>
<td></td>
</tr>
<tr>
<td>intra-</td>
<td>IOS</td>
</tr>
<tr>
<td>occipital</td>
<td></td>
</tr>
<tr>
<td>sulcus</td>
<td></td>
</tr>
<tr>
<td>intraparietal sulcus</td>
<td></td>
</tr>
</tbody>
</table>
ITG  inferior temporal gyrus
ITS  inferior temporal sulcus
LG   lingual gyrus
MFG  middle frontal gyrus
MOG  middle occipital gyrus
OG   orbital gyrus
poCG postcentral gyrus
po   pars opercularis of Broca’s area
poCS postcentral sulcus
POF  parieto-occipital fissure
preCS precentral sulcus
preCG precentral gyrus
precun precuneus
preSMA pre-supplementary motor area
pt   pars triangularis of Broca’s area
SFG  superior frontal gyrus
SFS  superior frontal sulcus
SMG  supramarginal gyrus
SOG  superior occipital gyrus
SPG  superior parietal gyrus
STS  superior temporal sulcus. All other abbreviations are as in the main text.

References


Foster NEV, Zatorre RJ. 2010; A role for the intraparietal sulcus in transforming musical pitch information. Cerebral Cortex. 20:1350–1359. [PubMed: 19789184]


Hutzler F. 2014; Reverse inference is not a fallacy per se: cognitive processes can be inferred from functional imaging data. NeuroImage. 84:1061–1069. [PubMed: 23313571]


Mastroberardino S, Santangelo V, Macaluso E. 2015 Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks.

Neuropsychologia. Author manuscript; available in PMC 2021 January 27.


Price CJ, Wise RJS, Frackowiak RSJ. 1996; Demonstrating the implicit processing of visually presented words and pseudowords. Cerebral Cortex. 6:62–70. [PubMed: 8670639]


Xie Y, Xu Y, Bian C, Li M. 2017; Semantic congruent audiovisual integration during the encoding stage of working memory: an ERP and sLORETA study. Scientific Reports. 7:5112.doi: 10.1038/s41598-017-05471-1 [PubMed: 28698594]

Highlights

Functional neuroimaging study of pitch-elevation crossmodal correspondence

Minimal evidence for mediation by semantic or multisensory integrative processes

Weak support for involvement of magnitude estimation

Possible relationship to phonological processing

Pitch-elevation congruency likely depends on multisensory attention/working memory
Figure 1.
Example stimuli for (a) the crossmodal pitch-elevation correspondence task and the localizer tasks: (b) semantic, (c) multisensory integration, (d) magnitude.
Figure 2.
Congruency effect from the CC > II contrast (pink) and the interaction (CC > II) > (IC > CI) contrast (blue).
Figure 3.
Congruency effect in relation to localizers. Pitch-elevation congruency effects (CC > II) – pink; semantic network (sentences > non-word lists) – orange; semantic control (non-word lists > sentences) – olive; multisensory integration (synchronous > asynchronous) – yellow; multisensory control (asynchronous > synchronous) – turquoise; magnitude network (magnitude > shape) – green. Circles indicate areas of overlap or contiguity between congruency effects and localizer conditions (see Table 4a).
Table 1

Semantic localizer activations: (a) complete sentences > non-words, (b) non-words > complete sentences, both within cortical mask, voxel-wise threshold $p < .001$, cluster-corrected $p < .05$, cluster threshold 8 voxels; $x,y,z$: Talairach coordinates for centers of gravity.

<table>
<thead>
<tr>
<th>Region</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R p STS - mid STS - a STS - a STG</td>
<td>49</td>
<td>−6</td>
<td>−9</td>
</tr>
<tr>
<td>R p calcS</td>
<td>14</td>
<td>−85</td>
<td>−2</td>
</tr>
<tr>
<td>L preSMA - SMA</td>
<td>−3</td>
<td>1</td>
<td>62</td>
</tr>
<tr>
<td>L preCS - preCG</td>
<td>−44</td>
<td>−5</td>
<td>47</td>
</tr>
<tr>
<td>L p STS - p STG - mid STS - a STS - a STG</td>
<td>−51</td>
<td>−14</td>
<td>−5</td>
</tr>
<tr>
<td>L pt</td>
<td>−45</td>
<td>27</td>
<td>−2</td>
</tr>
<tr>
<td>L p cingG</td>
<td>−4</td>
<td>−53</td>
<td>12</td>
</tr>
<tr>
<td>(b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R SFG</td>
<td>22</td>
<td>20</td>
<td>56</td>
</tr>
<tr>
<td>R med SFG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R MFG</td>
<td>33</td>
<td>19</td>
<td>46</td>
</tr>
<tr>
<td>R MFG - IFS</td>
<td>39</td>
<td>21</td>
<td>37</td>
</tr>
<tr>
<td>R IFG - IFS - MFG</td>
<td>41</td>
<td>40</td>
<td>16</td>
</tr>
<tr>
<td>R mid cingS - p cingG - p cingS</td>
<td>5</td>
<td>−26</td>
<td>37</td>
</tr>
<tr>
<td>R AG</td>
<td>49</td>
<td>−54</td>
<td>36</td>
</tr>
<tr>
<td>L p cingG - p cingS</td>
<td>−4</td>
<td>−26</td>
<td>28</td>
</tr>
<tr>
<td>L MFG - IFS</td>
<td>−37</td>
<td>26</td>
<td>36</td>
</tr>
<tr>
<td>L IFG - OG</td>
<td>−33</td>
<td>50</td>
<td>3</td>
</tr>
<tr>
<td>L POF - SPG</td>
<td>−13</td>
<td>−67</td>
<td>32</td>
</tr>
<tr>
<td>L med SFG</td>
<td>−3</td>
<td>30</td>
<td>32</td>
</tr>
<tr>
<td>L SMG - AG</td>
<td>−48</td>
<td>−48</td>
<td>41</td>
</tr>
<tr>
<td>L p Ins - mid Ins - po</td>
<td>−41</td>
<td>−6</td>
<td>6</td>
</tr>
</tbody>
</table>
Table 2
Multisensory integration localizer activations: (a) synchronous > asynchronous, (b) asynchronous > synchronous, both within cortical mask, voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 8 voxels; x,y,z: Talairach coordinates for centers of gravity.

<table>
<thead>
<tr>
<th>Region &amp; COG</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R SOG</td>
<td>6</td>
<td>-84</td>
<td>23</td>
</tr>
<tr>
<td>L POF - p cingG</td>
<td>-4</td>
<td>-59</td>
<td>15</td>
</tr>
<tr>
<td>(b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R AG - mid IPS - a IPS</td>
<td>36</td>
<td>-54</td>
<td>44</td>
</tr>
<tr>
<td>R MFG - IFS - IFG</td>
<td>44</td>
<td>16</td>
<td>34</td>
</tr>
<tr>
<td>L p IPS - a IPS</td>
<td>-37</td>
<td>-56</td>
<td>46</td>
</tr>
</tbody>
</table>
Table 3
Magnitude localizer activations: magnitude estimation > control within cortical mask, voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 7 voxels; x,y,z: Talairach coordinates for centers of gravity.

<table>
<thead>
<tr>
<th>Region</th>
<th>x</th>
<th>y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>R SMG - a IPS</td>
<td>42</td>
<td>-39</td>
<td>48</td>
</tr>
<tr>
<td>R SPG - av IPS</td>
<td>24</td>
<td>-61</td>
<td>40</td>
</tr>
<tr>
<td>R pv IPS</td>
<td>26</td>
<td>-68</td>
<td>34</td>
</tr>
<tr>
<td>R MOG</td>
<td>27</td>
<td>-87</td>
<td>7</td>
</tr>
</tbody>
</table>
Table 4
Pitch-elevation congruency-related activations: (a) CC > II, (b) interaction contrast (CC > II) > (IC > CI), both within cortical mask, voxel-wise threshold p < .001, cluster-corrected p < .05 cluster threshold (a) 7 voxels, (b) 6 voxels; x,y,z: Talairach coordinates for centers of gravity.

<table>
<thead>
<tr>
<th>Region</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R FEF</td>
<td>22</td>
<td>−6</td>
<td>57</td>
</tr>
<tr>
<td>R a IFG*</td>
<td>31</td>
<td>56</td>
<td>3</td>
</tr>
<tr>
<td>R FO</td>
<td>45</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>R mid Ins</td>
<td>40</td>
<td>−1</td>
<td>7</td>
</tr>
<tr>
<td>R a Ins</td>
<td>32</td>
<td>22</td>
<td>−1</td>
</tr>
<tr>
<td>R AG - mid IPS\textsuperscript{1,2,3,4}</td>
<td>47</td>
<td>−53</td>
<td>39</td>
</tr>
<tr>
<td>L po - mid Ins\textsuperscript{2}</td>
<td>−44</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>L a Ins\textsuperscript{5}</td>
<td>−32</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>(b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R AG - mid IPS</td>
<td>48</td>
<td>−55</td>
<td>40</td>
</tr>
<tr>
<td>R MFG</td>
<td>35</td>
<td>27</td>
<td>44</td>
</tr>
<tr>
<td>R a IFG</td>
<td>33</td>
<td>57</td>
<td>2</td>
</tr>
<tr>
<td>L SMG</td>
<td>−54</td>
<td>−51</td>
<td>35</td>
</tr>
<tr>
<td>L AG</td>
<td>−43</td>
<td>−54</td>
<td>34</td>
</tr>
<tr>
<td>L OG</td>
<td>−23</td>
<td>55</td>
<td>−4</td>
</tr>
</tbody>
</table>

\textsuperscript{1} Overlaps with the interaction result in (b).
\textsuperscript{2} Overlaps with the semantic control R AG/L po - mid Ins in Table 1b
\textsuperscript{3} Contiguous with the asynchronous R AG and close to R mid IPS in Table 3b.
\textsuperscript{4} Contiguous with the semantic control R mid IPS in Table 1b.
\textsuperscript{5} Close to the semantic L pt in Table 1a.