

2021

Meta-analyses support a taxonomic model for representations of different categories of audio-visual interaction events in the human brain

Matt Csonka
West Virginia University

Nadia Mardmomen
West Virginia University

Paula J. Webster
West Virginia University

Julie A. Brefczynski-Lewis
West Virginia University

Chris Frum
West Virginia University

See next page for additional authors

Follow this and additional works at: https://researchrepository.wvu.edu/faculty_publications



Part of the [Neuroscience and Neurobiology Commons](#)

Digital Commons Citation

Csonka, Matt; Mardmomen, Nadia; Webster, Paula J.; Brefczynski-Lewis, Julie A.; Frum, Chris; and Lewis, James W., "Meta-analyses support a taxonomic model for representations of different categories of audio-visual interaction events in the human brain" (2021). *Faculty & Staff Scholarship*. 2975.
https://researchrepository.wvu.edu/faculty_publications/2975

This Article is brought to you for free and open access by The Research Repository @ WVU. It has been accepted for inclusion in Faculty & Staff Scholarship by an authorized administrator of The Research Repository @ WVU. For more information, please contact ian.harmon@mail.wvu.edu.

Authors

Matt Csonka, Nadia Mardmomen, Paula J. Webster, Julie A. Brefczynski-Lewis, Chris Frum, and James W. Lewis

Meta-analyses support a taxonomic model for representations of different categories of audio-visual interaction events in the human brain

Matt Csonka
Nadia Mardmomen
Paula J. Webster
Julie A. Brefczynski-Lewis
Chris Frum
James W. Lewis

*Department of Neuroscience,
Rockefeller Neurosciences Institute
West Virginia University, Morgantown, WV 26506, USA*

Updated Dec 30, 2020

Key words: multisensory integration, categorical perception, sensory-semantic categories; embodied cognition, neuroimaging

Correspondence should be addressed to:

James W. Lewis, PhD.
Rockefeller Neurosciences Institute
Department of Neuroscience
P.O. Box 9303
West Virginia University
Morgantown, WV 26506
Phone: 304 293-1517
Fax: 304-293-3850
Email: jwlewis@hsc.wvu.edu
3 Figures (1 color), 14 Tables, Online Materials

Abstract

Our ability to perceive meaningful action events involving objects, people and other animate agents is characterized in part by an interplay of visual and auditory sensory processing and their cross-modal interactions. However, this multisensory ability can be altered or dysfunctional in some hearing and sighted individuals, and in some clinical populations. The present meta-analysis sought to test current hypotheses regarding neurobiological architectures that may mediate audio-visual multisensory processing. Reported coordinates from 82 neuroimaging studies (137 experiments) that revealed some form of audio-visual interaction in discrete brain regions were compiled, converted to a common coordinate space, and then organized along specific categorical dimensions to generate activation likelihood estimate (ALE) brain maps and various contrasts of those derived maps. The results revealed brain regions (cortical “hubs”) preferentially involved in multisensory processing along different stimulus category dimensions, including (1) living versus non-living audio-visual events, (2) audio-visual events involving vocalizations versus actions by living sources, (3) emotionally valent events, and (4) dynamic-visual versus static-visual audio-visual stimuli. These meta-analysis results are discussed in the context of neurocomputational theories of semantic knowledge representations and perception, and the brain volumes of interest are available for download to facilitate data interpretation for future neuroimaging studies.

Introduction

The perception of different categories of visual (unisensory) object and action forms are known to differentially engage distinct brain regions or networks in neurotypical individuals, such as when observing or identifying faces, body parts, living things, houses, fruits and vegetables, and outdoor scenes, among other proposed categories (Caramazza & Mahon, 2003; Martin, 2007; Martin, Wiggs, Ungerleider, & Haxby, 1996; Tranel, Damasio, & Damasio, 1997). Distinct semantic categories of real world unisensory sound-producing events are also known or thought to recruit different brain networks, such as nonliving environmental and mechanical sounds (Lewis, Talkington, Tallaksen, & Frum, 2012), non-vocal action events produced by non-human animal sources (Engel, Frum, Puce, Walker, & Lewis, 2009; Lewis, Talkington, Puce, Engel, & Frum, 2011), as well as the more commonly studied categories of living things (especially human conspecifics) and vocalizations (notably speech) (Brefczynski-Lewis & Lewis, 2017; Dick et al., 2007; Goll, Crutch, & Warren, 2011; Saygin, Leech, & Dick, 2010; Trumpp, Kliese, Hoenig, Haarmeier, & Kiefer, 2013). Extending beyond unisensory category-specific percepts, the neurobiological representations of multisensory events are thought to develop based on complex combinations of sensory and sensory-motor information, with some dependence on differences with individual observers' experiences throughout life, such as with handedness (Lewis, Phinney, Brefczynski-Lewis, & DeYoe, 2006). One may have varying experiences with, for instance, observing and hearing a construction worker hammering a nail, feeling a warm purring gray boots breed cat on a sofa. While watching television or a smart phone device one can readily accept the illusion that the synchronized audio (speakers) and video movements (the screen) are emanating from a single animate source, leading to stable,

unified multisensory percepts. Psychological literature indicates that perception of multisensory events can manifest as well-defined category-specific objects and action representations that build on past experiences (Martin, 2007; McClelland & Rogers, 2003; Miller, Nieder, Freedman, & Wallis, 2003; Rosch, 1973; Vygotsky, 1978).

However, the rules that may guide the organization of cortical network representations that mediate multisensory perception of real-world events, and whether any taxonomic organizations for such representations exist at a categorical level, remains unclear.

The ability to organize information to attain a sense of global coherence, meaningfulness, and possible intention behind every-day observable events may fail to fully or properly develop, as for some individuals with autism spectrum disorder (ASD) (Happe & Frith, 2006; Jolliffe & Baron-Cohen, 2000; Kouijzer, de Moor, Gerrits, Congedo, & van Schie, 2009; Marco, Hinkley, Hill, & Nagarajan, 2011; B. Pfeiffer, Clark, & Arbesman, 2018; B. A. Pfeiffer, Koenig, Kinnealey, Sheppard, & Henderson, 2011; Powers, Hillock, & Wallace, 2009; Ramot et al., 2017; Webster et al., 2020) and possibly for some individuals with various forms of schizophrenia (Cecere, Gross, & Thut, 2016; Roa Romero, Keil, Balz, Gallinat, & Senkowski, 2016; Straube, Green, Sass, & Kircher, 2014; Vanes et al., 2016). Additionally, brain damage, such as with stroke, has been reported to lead to deficits in multisensory processing (Van der Stoep, Van der Stigchel, Van Engelen, Biesbroek, & Nijboer, 2019). Thus, further understanding the organization of the multisensory brain has been becoming a topic of increasing clinical relevance.

At some processing stages or levels, the central nervous system is presumably “pre-wired” to readily develop an organized architecture that can rapidly and efficiently

extract meaningfulness from multisensory events. This includes audio-visual event encoding and decoding that enables a deeper understanding of one's environment, thereby conferring a survival advantage through improvements in perceived threat detection and in social communication (Donald, 1991; Hewes, 1973; Rilling, 2008; Robertson & Baron-Cohen, 2017). An understanding of multisensory neuronal processing mechanisms, however, may in many ways be better understood through models of semantic knowledge processing rather than models of bottom-up signal processing, which is prevalent in unisensory fields of literature. One set of theories behind semantic knowledge representation includes distributed-only views, wherein auditory, visual, tactile and other sensory-semantic systems are distributed neuroanatomically with additional task dependent representations or convergence-zones in cortex that link knowledge (A.R. Damasio, 1989; H. Damasio, Grabowski, Tranel, Hichwa, & Damasio, 1996; Ghazanfar & Schroeder, 2006; Languis & Miller, 1992; Martin, 2007; Tranel et al., 1997). A distributed-plus-hub view further posits the existence of additional task-independent representations (or "hubs") that support the interactive activation of representations in all modalities, and for all semantic categories (Patterson, Nestor, & Rogers, 2007).

More recent neurocomputational theories of semantic knowledge learning entails a sensory-motor framework wherein action perception circuits (APCs) are formed through sensory experiences, which manifest as specific distributions across cortical areas (Pulvermuller, 2013, 2018; Tomasello, Garagnani, Wennekers, & Pulvermuller, 2017). In this construct, combinatorial knowledge is thought to become organized by connections and dynamics between APCs, and cognitive processes can be modelled forthright. Such

models have helped to account for the common observation of cortical hubs or ‘connector hubs’ for semantic processing (A. R. Damasio, 1989; Sporns, Honey, & Kotter, 2007; van den Heuvel & Sporns, 2013), which may represent multimodal, supramodal, or amodal mechanisms for representing knowledge. From this connector hub theoretical perspective, it remains unclear whether or how different semantic categories of multisensory perceptual knowledge might be organized, potentially including semantic hubs that link, for instance, auditory and visual unisensory systems at a category level.

Here we addressed the issue of global neuronal organizations that mediate different aspects of audio-visual categorical perception using activation likelihood estimate (ALE) meta-analyses of a diverse range of published studies to date that reported audio-visual interactions of some sort in the human brain. We defined the term “interaction” to include measures of neuronal sensitivity to temporal and/or spatial correspondence, response facilitation or suppression, inverse effectiveness, an explicit comparison of information from different modalities that pertained to a distinct object, and cross-modal priming (Calvert & Lewis, 2004; Stein & Meredith, 1990; Stein & Wallace, 1996). These interaction effects were assessed in neurotypical adults (predominantly, if not exclusively, right-handed) using hemodynamic blood flow measures [functional magnetic resonance imaging (fMRI), or positron emission tomography (PET)] or magnetoencephalography (MEG) methodologies as whole brain neuroimaging techniques.

The resulting descriptive compilations and analytic contrasts of audio-visual interaction sites across different categories of audio-visual stimuli were intended to meet three main goals: The first goal was to reveal a global set of brain regions (cortical and

non-cortical) with significantly high probability of cross-sensory interaction processing regardless of variations in methods, stimuli, tasks, and experimental paradigms. The second goal was to validate and refine earlier multisensory processing concepts borne out of image-based meta-analyses of audio-visual interaction sites (Lewis, 2010) that used a subset of the paradigms included in the present study, but here taking advantage of coordinate-based meta-analyses and more rigorous statistical approaches now that additional audio-visual interaction studies have subsequently been published.

The third goal, as a special focus, was to test recent hypotheses regarding putative brain architectures mediating multisensory categorical perception that were derived from unisensory auditory object perception literature (Fig. 1), which encompassed theories to explain how real-world natural sounds are processed to be perceived as meaningful events to the observer (Brefczynski-Lewis & Lewis, 2017). This hearing perception model entailed four proposed tenets that may shape brain organizations for processing real-world natural sounds, helping to explain *why* certain category-preferential representations appear in the human brain (and perhaps more generally in the brains of all mammals with hearing ability). These tenets for hearing perception included: (A) parallel hierarchical pathways process increasing information content, (B) metamodal operators guide sensory and multisensory processing network organizations, (C) natural sounds are embodied when possible, and (D) categorical perception emerges in neurotypical listeners.

Figure 1 near here

After compiling the numerous multisensory human neuroimaging studies that employed different types of audio-visual stimuli, tasks, and imaging modalities, we

sought to test three hypotheses relating to the above mentioned tenets and neurobiological model. The first two hypotheses effectively tested for support of the major taxonomic boundaries depicted in Figure 1: The first hypothesis being (1) that there will be a double-dissociation of brain systems for processing living versus non-living audio-visual events, and the second hypothesis (2) that there will be a double-dissociation for processing vocalizations versus action audio-visual events produced by living things. In the course of compiling neuroimaging literature, there was a clear divide between studies using static visual images (iconic representations) versus video with dynamic motion stimuli that corresponded with aspects of the auditory stimuli. The production of sound necessarily implies dynamic motion of some sort, which in many of the studies' experimental paradigms also correlated with viewable object or agent movements. Thus, temporal and/or spatial intermodal invariant cues that physically correlate visual motion ("dynamic-visual") with changes in acoustic energy are typically uniquely present in experimental paradigms using video (Bulkin & Groh, 2006; Lewkowicz, 2000; Stein & Meredith, 1993). Conversely, static or iconic visual stimuli ("static-visual") must be learned to be associated and semantically congruent with characteristic sounds, and with varying degrees of arbitrariness. Thus, a third hypothesis emerged (3) that the processing of audio-visual stimuli that entailed dynamic-visual motion stimuli versus static-visual stimuli will also reveal a double-dissociation of cortical processing pathways in the multisensory brain. The identification and characterization of any of these hypothesized neurobiological processing categories at a meta-analysis level would newly inform neurocognitive theories, specifying regions or network hubs where certain types of information may merge or in some way interact across sensory systems at a semantic

category level. Thus, the resulting ALE maps are expected to facilitate the generation of new hypotheses regarding multisensory interaction and integration mechanisms in neurotypical individuals. They should also contribute to providing a foundation for ultimately understanding *why* multisensory processing networks develop the way they typically do, and why they may develop aberrantly, or fail to recover after brain injury, in certain clinical populations.

Methods

This work was performed in accordance with the PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions (Moher, Liberati, Tetzlaff, Altman, & Group, 2009). Depicted in the PRISMA flow-chart (Fig. 2), original research studies were identified by PubMed and Google Scholar literature searches with keyword combinations “auditory + visual”, “audiovisual”, “multisensory”, and “fMRI” or “PET” or “MEG”, supplemented through studies identified through knowledge of the field published between 1999 through early 2020. Studies involving drug manipulations, patient populations, children, or non-human primates were excluded unless there was a neurotypical adult control group with separately reported outcomes. Of the included studies, reported coordinates for some paradigms had to be estimated from figures. Additionally, some studies did not use whole-brain imaging, but rather incorporated imaging to a 50 to 60 mm slabs of axial brain slices so as to focus, for instance, on the thalamus or basal ganglia. These studies were included despite their being a potential violation of assumptions made by ALE analyses (see below) because the emphasis of the present study was to reveal proof of

concept regarding differential audio-visual processing at a semantic category level. This yielded inclusion of 82 published fMRI, PET and MEG studies including audio-visual interaction(s) of some form (Table 1). The compiled coordinates, after converting to afni-TLRC coordinate system, derived from these studies are included in Appendix A, and correspond directly to Table 1.

Figure 2 near here

Table 1 near here

Activation Likelihood Estimate (ALE) Analyses

The ALE analysis consists of a coordinate-based, probabilistic meta-analytic technique for assessing the co-localization of reported activations across studies (Eickhoff, Bzdok, Laird, Kurth, & Fox, 2012; Eickhoff et al., 2009; Eickhoff et al., 2016; A. R. Laird et al., 2009; N. Laird, Fitzmaurice, & Ding, 2010; Muller et al., 2018; Turkeltaub, Eden, Jones, & Zeffiro, 2002; Turkeltaub et al., 2012). Whole-brain probability maps were initially created across all the reported foci in standardized stereotaxic space (Talairach “T88”, being converted from, for example, Montreal Neurological Institute “MNI” format) using GingerALE software (Brainmap GingerALE version 2.3.6; Research Imaging Institute; <http://brainmap.org>). This software was also used to create probability maps, where probabilities were modeled by 3D Gaussian density distributions that took into account sample size variability by adjusting the full-width half-max (FWHM) for each study (Eickhoff et al., 2009; Eickhoff et al., 2016). For each voxel, GingerALE estimated the cumulative probabilities that at least one study reported activation for that locus for a given experimental paradigm condition. Assuming and accounting for spatial uncertainty across reports, this voxel-wise procedure generated

statistically thresholded ALE maps, wherein the resulting ALE values reflected the probability of reported activation at that locus. Using a random effects model, the values were tested against the null hypothesis that activation was independently distributed across all studies in a given meta-analysis.

To determine the likely spatial convergence of reported activations across studies, activation foci coordinates from experimental paradigms were transferred manually and compiled into one spreadsheet on two separate occasions by two different investigators (co-authors). To avoid (or minimize) the potential for errors (e.g. transformation from MNI to TAL, sign errors, duplicates, omissions, etc.) an intermediate stage of data entry involved logging all the coordinates and their transformations into one spreadsheet (Appendix A) where they were coded by Table/Figure and number of subjects (Table 1), facilitating inspection and verification relative to hard copy printouts of all included studies. A third set of files (text files) were then constructed from that spreadsheet of coordinates and entered as input files for the various meta-analyses using GingerALE software. This process enabled a check-sums of number of left and right hemisphere foci and the number of subjects for all of the meta-analyses reported herein. When creating single study data set analysis ALE maps, coordinates from experimental paradigms of a given study (using the same participants in each paradigm) were pooled together, thereby avoiding potential violations of assumed subject-independence across maps, which could negatively impact the validity of the meta-analytic results (Turkeltaub et al., 2012). After pooling, there were 1,285 participants (Table 1, column 6). Some participants could conceivably have been recruited in more than one study (such as from the same laboratory). However, we had no means for assessing this and assumed that these were all

unique individuals. All single study data set ALE maps were thresholded at $p < .05$ and using a voxel-level family-wise error (FWE) rate correction for multiple comparisons (Muller et al., 2018) using 10,000 Monte Carlo threshold permutations. For all “contrast” ALE meta-analysis maps, cluster-level thresholds were derived using the single study corrected FWE datasets and then further thresholded for contrast at an uncorrected $p < .05$, and using 10,000 permutations. Minimum cluster sizes were used to further assess rigorosity of clusters, which are included in the tables and addressed in the results section.

Guided by earlier meta-analyses of hearing perception and audio-visual interaction sites several hypothesis driven contrasts were derived, as addressed in the Introduction (Brefczynski-Lewis & Lewis, 2017; Lewis, 2010). A minimum of 17-20 studies was generally recommended to achieve sufficient statistical power to detect smaller effects and make sure that results were not driven by single experiments (Eickhoff et al., 2016; Muller et al., 2018). However, two of the ten sub-sets of meta-analysis were performed despite there being relatively few numbers of studies (i.e., $n=13$ in Table 9; $n=9$ in Table 10), and thus their outcomes would presumably only reveal the larger effect sizes. For visualization purposes, resulting maps were initially projected onto the N27 atlas brain using AFNI software (Cox, 1996) to assess and interpret results, and onto the Population-Averaged, Landmark- and Surface-based (PALS) atlas cortical surface models (in AFNI-Talairach space) using Caret software (<http://brainmap.wustl.edu>) for illustration of the main findings (Van Essen, 2005; Van Essen et al., 2001).

Results

The database search for audio-visual experiments reporting interaction effects yielded 137 experimental paradigms from 82 published articles (Fig. 2; PRISMA flow-chart). Experiments revealing an effect of audio-visual stimuli (Table 1) included 1,285 subjects (though see Methods) and 714 coordinate brain locations (376 left hemisphere, 338 right). ALE meta-analysis of all these reported foci (Congruent plus Incongruent audio-visual interaction effects) revealed a substantial expanse of activated brain regions (Fig. 3A, purple hues; projected onto both fiducial and inflated brain model images). Note that this unthresholded map revealed foci reported as demonstrating audio-visual interactions that were found to be significant in at least one of the original studies, thereby illustrating the substantial global expanse of reported brain territories involved in audio-visual interaction processing in general. This included sub-cortical in addition to cortical regions, such as the thalamus and basal ganglia (Fig. 3A insets), and cerebellum (not illustrated). However, sub-cortical regions are only approximately illustrated here since they did not survive threshold criteria imposed in the below single study and contrast ALE brain maps. Each study contained one or multiple experimental paradigms. For each experimental paradigm, several neurobiological sub-categories of audio, visual and/or audio-visual stimuli were identified. The sub-categories are coded in Table 1 (far right columns) as either being excluded (0), included (1), included as a contrast condition (2), or deemed as uncertain for inclusion (blank cells) for use in different meta-analyses, which are detailed by section below. Volumes resulting from the meta-analyses (depicted in Fig. 3) are available for download as Online Supplemental Data.

Figure 3 near here

Congruent versus incongruent audio-visual stimuli. The first set of meta-analyses examined reported activation foci specific to when audio-visual stimuli were perceived as congruent spatially, temporally and/or semantically (Table 2; 79 studies, 117 experimental paradigms, 1,235 subjects, 608 reported foci—see Table captions) versus those regions more strongly activated when the stimuli were perceived as incongruent (Table 3). Brain maps revealing activation when processing only congruent audio-visual pairings (Congruent single study; corrected FWE $p < .05$) revealed several regions of interest (ROIs) (Fig. 3B, white hues; Table 4A coordinates), including the bilateral posterior superior temporal sulci (pSTS) that extended into the bilateral planum temporale and transverse temporal gyri (left > right), and the bilateral inferior frontal cortices (IFC). Brain maps revealing activation when processing incongruent audio-visual pairings (Fig. 3B, black hues; Incongruent single study; corrected FWE $p < .05$; Table 4B) revealed bilateral IFC foci that were located immediately anterior to the IFC foci for congruent stimuli, plus a small left anterior insula focus.

A contrast meta-analysis of Congruent > Incongruent audio-visual stimuli (Fig. 3B, white with black outlines; Table 4C, uncorrected $p < .05$) revealed significant involvement of the left and right posterior temporal gyri (pSTG) and pSTS regions. Conversely, a contrast map of brain regions showing significant preferential involvement in processing Incongruent > Congruent audio-visual stimuli (Fig. 3B, black with white outlines; Table 4D, uncorrected $p < .05$) included bilateral inferior frontal cortices (IFC), which extended along inferior portions of the middle frontal gyri in locations immediately anterior to those resulting from the Congruent > Incongruent contrast. Because both contrast ALE maps revealed functionally dissociated regions of interest

(ROIs), these results are herein regarded as providing evidence for a ‘double-dissociation’ of processing along this dimension.

Tables 2, 3 and 4 near here

Living versus non-living audio-visual stimuli. A major categorical distinction in the neurobiological organization mediating auditory perception is that for sounds produced by living versus non-living sources (Fig. 1). This potential categorical processing boundary was tested in the multisensory realm by comparing reported activation foci from audio-visual interaction paradigms that involved Living versus Non-living sources. The Living category paradigms included visual and/or sound-source stimuli such as talking faces, hand/arm gesture with speech, body movements, tool use, and non-human animals (Table 5; see brief descriptions). A single study ALE meta-analysis of experimental paradigms using Living stimuli revealed portions of the bilateral pSTS/pSTG regions (Fig. 3C, orange hues; Table 7A, corrected FWE $p < .05$). The Non-living visual and sound-source stimuli (Table 6) predominantly included artificial, as opposed to natural, audio-visual events such as flashing checkerboards, coherent dot motion, geometric objects (plus a study depicting natural environmental events), which were paired with sounds such as tones, sirens, or mechanical sounds produced by inanimate sources. A single study ALE meta-analysis of experiments using Non-living stimuli (mostly artificial stimuli) revealed the right anterior insula as a region significantly recruited (Fig. 3C, cyan contained within the white outline; Table 7B, corrected FWE $p < .05$; Also see contrast below).

A contrast ALE meta-analysis of maps Living > Non-living events revealed bilateral pSTS foci as showing significant differential responsiveness (Fig. 3C, orange with

outline [visible only in left hemisphere model]; Table 7C, uncorrected $p < .05$). The contrast meta-analysis of Non-living > Living congruent audio-visual events revealed the right anterior insula as a common hub of activation (Fig. 2C, cyan with white outline; Table 7D, uncorrected $p < .05$). A main contributing study to this right anterior insula ROI (study #44 Meyer et al., 2007) included screen flashes paired with phone rings as part of a conditioned learning paradigm.

In visual perception literature, a prominent dichotomy of stimulus processing involves “what versus where” streams (Goodale et al., 1994; L. G. Ungerleider & Haxby, 1994; L.G. Ungerleider, Mishkin, Goodale, & Mansfield, 1982), which has also been explored in the auditory system (Clarke et al., 2002; Kaas & Hackett, 1999; J. P. Rauschecker, 1998; J. P. Rauschecker & Scott, 2015; J.P. Rauschecker & Tian, 2000). A few of the audio-visual interaction studies examined in the present meta-analyses either explicitly or implicitly tested that organization (Plank, Rosengarth, Song, Ellermeier, & Greenlee, 2012; Sestieri et al., 2006). However there were insufficient numbers of studies germane to that dichotomy for conducting a proper meta-analysis along this dimension.

Tables 5, 6 and 7 near here

Vocalization versus Action event audio-visual interaction sites. Another stimulus category boundary derived from auditory categorical perception literature was that for processing vocalizations versus action sounds (Fig. 1). To be consistent with that neurobiological model, this category boundary was tested using only *living* audio-visual sources. This analysis included vocalizations by human or animal sources (Table 8) versus action events (Table 9) including sounds produced by, for example, hand tool use, bodily actions, and persons playing musical instruments. An ALE single study map for

experiments using Vocalizations revealed four ROIs along the pSTG/pSTS region (Fig. 3D, red hues; Table 11A, corrected FWE $p < .05$). The action event category was initially restricted to using only non-vocalizations (by living things) as auditory stimuli. This initially yielded 9 studies that showed audio-visual interaction foci, and no clusters survived the single study ALE meta-analysis map voxel-wise thresholding. Adjusting the study restrictions to include studies that reported using a mix of action events together with some non-living visual stimuli and some vocalizations as auditory (non-verbal) event stimuli yielded 13 studies (Table 9). A single study ALE map for these Action events, which were predominantly non-vocal and depicting living things, revealed one ROI along the left fusiform gyrus (Table 11B, corrected FWE $p < .05$).

The contrast meta-analysis of Vocalizations > Actions revealed right pSTS and pSTG foci as being preferential for vocalizations (Fig. 3D, red with black outlines; Table 11C, uncorrected $p < .05$). Conversely, the contrast meta-analysis of Action > Vocalization audio-visual interactions revealed the left fusiform gyrus ROIs (Fig. 3D, yellow with black outline; Table 11D, uncorrected $p < .05$). This left fusiform ROI had a volume of 8 mm³, both in the single study and contrast ALE meta-analysis maps. This cluster size fell below some criteria for rigor depending on theoretical interpretation when group differences are diffuse (Tench, Tanasescu, Auer, Cottam, & Constantinescu, 2014). Nonetheless, this theoretical processing dissociation existed in at least some single studies, in the single ALE map, and in the contrast ALE map, and was thus at least suggestive of a double-dissociation. A main contributing study to this fusiform ROI (study #62, Schmid et al., 2011) employed a relatively simple task of determining if a colored picture included a match to a presented sound, or vice versa, which involved a

wide variety of non-living but a few living real-world object images. This ROI was consistent in location with the commonly reported fusiform foci involved in functions pertaining to high-level visual object processing (Bar et al., 2001; Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999; James & Gauthier, 2003).

A subset of the paradigms involving Living things and/or Vocalizations included emotionally valent stimuli (Table 10). This predominantly including emotional faces with voice (expressing fear, anger, sadness, happiness and laughter), but also whole body and dance expressions rated for emotional content. These emotionally valent paradigms preferentially activated a portion of the right pSTG (Fig. 3D, violet hues; Table 11E), when analyzed as a single study ALE map (corrected FWE $p < .05$), but also as a contrast meta-analysis with non-emotionally valent paradigms involving living things (mostly control conditions from the same or similar paradigms; data not shown).

Tables 8, 9, 10 and 11 near here

Dynamic visual motion versus static images in audio-visual interactions. We next sought to determine if the use of dynamic visual motion versus static visual images in audio-visual interaction paradigms might reveal differences in processing organizations in the brain. Studies using Dynamic-visual stimuli (Table 12), included talking faces, the McGurk effect, hand gestures, bodily gestures, and geometric shapes modulating synchronously with vocals, plus non-vocal drum sounds, musical instruments (e.g. piano), hand tool sounds, tone sweeps, and synthetic tones. Studies using Static-visual images (Table 13) involved the matching of pictures of human faces or animals to characteristically associated vocal sounds, plus other forms of photos or drawings (in color, grayscale, or black and white) of faces, animals, objects, or written word/character

forms, while excluding stimuli such as flashing screens or light emitting diodes (LEDs). ALE single study maps for experiments utilizing Dynamic-visual stimuli (Fig. 3E, blue hues; Table 14A, corrected FWE $p < .05$) and Static-visual stimuli (pink hues; Table 14B, corrected FWE $p < .05$) were constructed. A contrast ALE meta-analysis of Dynamic-visual $>$ Static-visual revealed significantly greater activation of the right pSTS region (Fig. 3E, blue with black outline; Table 14C, uncorrected $p < .05$). Conversely, the contrast ALE meta-analysis of Static-visual $>$ Dynamic-visual paradigms preferentially activated the bilateral planum temporale and STG regions (Fig. 3E, pink with black outlines; Table 14D, uncorrected $p < .05$).

Analyses of the Dynamic-visual vs Static-visual was further conducted separately for those experimental paradigms using artificial versus natural stimuli. With the exception of natural Dynamic-visual studies ($n=37$ of the 43 in Table 12) the other sub-categories had too few studies for the recommended 17-20 study minimum for meta-analysis. Nonetheless, the artificial Static-visual ($n=12$) meta-analysis revealed clusters that overlapped with the outcomes using the respective full complement of studies, while the natural Dynamic-visual ($n=37$) meta-analysis revealed clusters that overlapped with the respective full complement of studies. Thus, audio-visual events involving dynamic visual motion (and mostly natural stimuli) generally recruited association cortices situated roughly between auditory and visual cortices, while audio-visual interactions involving static (iconic) visual images (and mostly artificial stimuli) generally recruited regions located closer to auditory cortex *proper* along the pSTG and planum temporale bilaterally.

Tables 12, 13 and 14 near here

Discussion

The present meta-analyses examined a wide variety of published human neuroimaging studies that revealed some form of audio-visual “*interaction*” in the brain, entailing responses beyond or different from the corresponding unisensory auditory and/or visual stimuli alone. One objective was to test several tenets regarding potential brain organizations or architectures that may develop for processing different categories of audio-visual event types at a semantic level. The tenets were borne out of recent ethologically derived unisensory hearing perception literature (Brefczynski-Lewis & Lewis, 2017). This included a taxonomic model of semantic categories of natural sound-producing events (i.e. Fig. 1), but here being applied to testing specific hypotheses in the realm of multisensory (audio-visual) processing. The category constructs were derived with the idea of identifying putative cortical “hubs” that could be further applied to, and tested by, various neurocomputational models of semantic knowledge and multisensory processing.

Providing modest support for our first hypothesis, contrast ALE meta-analyses revealed a double-dissociation of brain regions preferential for the processing of living versus non-living (mostly artificial sources) audio-visual interaction events at a category level (Fig. 3C, orange vs cyan). These results implicated the bilateral pSTS complexes versus the right anterior insula as processing hubs, respectively, which are further addressed below in Section A. Providing modest support for our second hypothesis, contrast ALE meta-analyses revealed a double-dissociation of brain regions preferential for the processing of audio-visual interaction events involving vocalizations versus actions, respectively (Fig. 3D, red vs yellow). These results implicated the bilateral

planum temporale, pSTG, and pSTS complexes versus the left fusiform cortex, respectively, which is also further addressed in Section A below.

Providing strong support for our third hypothesis, different cortices were preferential for processing audio-visual interactions that involved dynamic-visual (video) versus static-visual (iconic images) as visual stimuli (Fig. 3E, blue vs pink). This finding is addressed further below in Section B in the context of parallel multisensory processing hierarchies. The original volumes of the regions of interest (ROIs) identified herein (comprising clusters in Tables 4, 7, 11 and 14, and depicted in Fig. 3) are available for download as Online Supplemental Data. These ROI volumes should facilitate the generation and testing of new hypotheses, especially as they pertain to neurocomputational theories of semantic knowledge representation, which is a topic addressed in Section C. This is followed by Section D that considers various limitations of the meta-analysis studies.

Upon inspection of Figure 3C-E, only ventral cortices, as opposed to dorsal cortices (e.g. superior to the lateral sulcus), revealed activation foci that were preferential for any of the different semantic categories of audio-visual events. In particular, neither the bilateral IFC foci for congruent versus incongruent audio-visual interactions (Fig. 3B, black/white), nor the frontal or parietal cortices (Fig. 3A, purple), revealed any differential activation along the semantic category dimensions tested. This was generally consistent with the classic ventral “what is it” (perceptual identification of objects) versus dorsal “where is it” (sensory transformations for guided actions directed at objects) dichotomy observed in both vision and auditory neuroimaging and primate literature (Belin & Zatorre, 2000; Goodale & Milner, 1992; Sestieri et al., 2006; L. G. Ungerleider

& Haxby, 1994). While dorsal cortical regions such as a bilateral parietal cortices and non-cortical regions such as the cerebellum were reported to be revealing audio-visual interaction effects by many studies, their involvement appeared to relate more to task demands and task difficulty rather than semantic category of the audio-visual events *per se*. Dorsal cortical networks are often implicated in various components of attention. While some form of sensory attention was involved in nearly all of the experimental paradigms, the specific effects of different types or degrees of sensory attention was not a measurable dimension across the studies, and thus fell outside the scope of the present study.

A. Embodied representations of audio-visual events. One of the tenets regarding the taxonomic category model of real-world hearing perception was that “*natural sounds are embodied when possible*” (Brefczynski-Lewis & Lewis, 2017), and this tenet appears to also apply to the context of cortical organizations for processing audio-visual interactions at a semantic category level. This is further addressed below by region in the context of (A.1) the pSTS complexes for embodiability representations, and of (A.2) the right anterior insula focus for non-embodiable non-living and artificial audio-visual event perception.

A.1. The pSTS complexes and audio-visual motion processing. The bilateral pSTS complexes were significantly more involved with processing audio-visual interactions associated with events by living things, by stimuli involving vocalizations, and by dynamic-visual (versus static-visual image) audio-visual events (cf. Fig. 3C-E, orange, red and blue). Although these respective foci were derived by several overlapping

studies, the meta-analysis results support the notion that the lateral temporal cortices are the primary loci for complex natural motion processing (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp, Lee, Argall, & Martin, 2004; Calvert, Campbell, & Brammer, 2000; Calvert & Lewis, 2004; Lewis et al., 2004; Martin, 2007; Taylor, Moss, Stamatakis, & Tyler, 2006; Taylor, Stamatakis, & Tyler, 2009): More specifically, the pSTS complexes are thought to play a prominent perceptual role in transforming the spatially and temporally dynamic features of natural auditory and visual action information together into a common neural code, which may then facilitate cross-modal interactions and integration from a “bottom-up” intermodal invariant sensory perspective. An earlier image-based (as opposed to coordinate-based) meta-analysis using a subset of these paradigms (Lewis, 2010) further highlighted the idea that the pSTS complexes may form a temporal reference frame for probabilistically comparing the predicted or expected incoming auditory and/or visual information based on what actions have already occurred.

From a “top-down” cognitive perspective, however, words and phrases that depict human actions, and even imagining complex audio and/or visual actions, are reported to lead to activation of the pSTS regions (Kellenbach, Brett, & Patterson, 2003; Kiefer, Sim, Herrnberger, Grothe, & Hoenig, 2008; Noppeney, Josephs, Hocking, Price, & Friston, 2008; Tettamanti et al., 2005). Furthermore, the pSTS complexes are known to be recruited by a variety of sensory-perceptual tasks in congenitally blind and in congenitally deaf individuals (Amedi et al., 2007; Burton, Snyder, & Raichle, 2004; Capek et al., 2008; Capek et al., 2010; Lewis, Frum, et al., 2011; Patterson et al., 2007; Pietrini et al., 2004), suggesting that aspects of their basic functional roles may not be

dependent on bimodal sensory input outright. To reconcile these findings, one hypothesis was that some cortical regions may develop to perform amodal or metamodal operations (Pascual-Leone & Hamilton, 2001). More specifically, different patches of cortex, such as the pSTS, may innately develop to contain circuitry predisposed to compete for the ability to perform particular types of operations or computations useful to the observer regardless of the modality (or presence) of sensory input. Thus, the organization of the multisensory brain may be influenced as much, if not more, by internal processing factors than by specific external sensory experiences *per se*. This interpretation reflects another tenet regarding the taxonomic category model of real-world hearing perception that “*metamodal operators guide sound processing network organizations*” (Brefczynski-Lewis & Lewis, 2017), but here applying to the processing of audio-visual interactions at a semantic category level.

Another interpretation regarding the functions of the bilateral pSTS complexes is that they are more heavily recruited by living and dynamic audio-visual events simply because of their greater life-long familiarity with adult observers. They may reflect an individual’s experiences and habits of extracting subtle nuances from day-to-day real-world interactions, including other orally communicating people as prevalent sources of multisensory events. Ostensibly, this experiential multisensory process would start from the time of birth when there becomes a critical need to interact with human caretakers. Consistent with this interpretation is that the pSTS complexes have prominent roles in social cognition, wherein reading subtleties of human expressions and body language is often highly relevant for conveying information that guides effective social interactions (Jellema & Perrett, 2006; Pelphrey, Morris, & McCarthy, 2004; Zilbovicius et al., 2006).

Embodied cognition models (also called grounded cognition) posit that perception of natural events (social or otherwise) is at least in part dependent on modal simulations, bodily states, and situated actions (Barsalou, 2008). The discovery of mirror neuron systems (MNS) and echo-mirror neuron (ENS) systems (Molenberghs, Cunnington, & Mattingley, 2012; Rizzolatti & Arbib, 1998; Rizzolatti & Craighero, 2004) have been recognized as having major implications for explaining many cognitive functions, including action understanding, imitation and empathy. Such neuronal systems, which often include the bilateral pSTS complexes, are proposed to mediate elements of the perception of sensory events as they relate to one's own repertoire of dynamic visual action-producing and sound-producing motoric events (Engel et al., 2009; Galati et al., 2008; Gazzola, Aziz-Zadeh, & Keysers, 2006; Lahav, Saltzman, & Schlaug, 2007; Lewis, Silberman, Donai, Frum, & Brefczynski-Lewis, 2018). Thus, the pSTS complexes may reflect metamodal cortices that typically develop to process natural multisensory events, which especially include dynamic actions by living things (including vocalizations) that are interpreted for meaningfulness (and possibly intent) based on embodiment strategies by the brain.

Notwithstanding, the dynamic viewable motions and sounds produced by non-living things and artificial stimulus events are arguably less embodyable or mimicable than those by living things. The pSTG/pSTS complexes were not preferentially activated by non-living and artificial multisensory events. Rather, this event category preferentially recruited the right anterior insula, as addressed next.

A.2. The right anterior insula and non-living/artificial audio-visual interaction

processing. The right anterior insula emerged as a cortical hub that was preferentially involved in processing congruent audio-visual events that entailed non-living and largely artificial sources, all of which were typically non-embodiable. Moreover, unlike the pSTS complexes, the right anterior insula did not show significant sensitivity to the dynamic-visual versus static-visual image stimulus dimension, suggesting that intermodal invariant cues were not a major driving factor in its recruitment. Interestingly, the mirror opposite left anterior insula showed preferential activation for Incongruent versus Congruent audio-visual stimuli (cf. Fig. 3B and 3C).

On a technical note, portions of the claustrum are located very close to the anterior insulae, and activation of the claustrum may have contributed to the anterior insula foci in several neuroimaging studies, and thus also in this meta-analysis. The enigmatic claustrum is reported to have a role in integrative processes that require the analysis of the *content* of the stimuli, and in coordinating the rapid integration of object attributes across different modalities that lead to coherent conscious percepts (Crick & Koch, 2005; Naghavi, Eriksson, Larsson, & Nyberg, 2007).

Embodiment encoding functions have been ascribed to the anterior insula in representing “self” versus “non-self”. For instance, the anterior insulae, which receive input from numerous cortical areas, have reported roles in multimodal integrative functions, re-representation of interoceptive awareness of bodily states, cognitive functions, and meta-cognitive functions (Craig, 2009, 2010; Menon & Uddin, 2010), and in social emotions that may function to help establish “other-related” states (Lamm & Singer, 2010; Singer et al., 2004). The right lateralized anterior insula activation has

further been reported to be recruited during non-verbal empathy-related processing such as with compassion meditation, which places an emphasis on dissolving the “self-vs-other” boundary (Lutz, Brefczynski-Lewis, Johnstone, & Davidson, 2008). Moreover, dysfunction of the anterior insulae has been correlated with an inability to differentiate the self from the non-self in patients with schizophrenia (Casella, Gerner, Fieldstone, Sawa, & Schretlen, 2011; Shura, Hurley, & Taber, 2014).

Although the anterior insula territories are commonly associated with affective states, visceral responses, and the processing of feelings (Cacioppo, 2013; Critchley, Wiens, Rotshtein, Ohman, & Dolan, 2004; Dalgleish, 2004; A. Damasio, 2001; Mutschler et al., 2009), the emotionally valent paradigms in this meta-analysis did not yield significant differential audio-visual interaction effects in the right (or left) insula, but rather only along the right pSTG. Though speculative, the anterior insula(e) may be subserving the mapping of events that are heightened by relatively ‘non-embodiable’ multisensory events (notably non-living and artificial sources) with differential activation depending on the perceived relatedness to self. This outcome will likely be a topic of interest for future studies, including neurocomputational modeling of cognition, which is addressed in a later section after first considering parallel multisensory processing hierarchies.

B. Dynamic-visual versus static-visual images and audio-visual interaction processing. The double-dissociation of cortical hubs for processing dynamic-visual versus static-visual audio-visual interactions was consistent with notion of parallel processing hierarchies. The experimental paradigms using video typically included dynamic intermodal invariant cross-modal cues (mostly by living things), where the

audio and visual stimuli were either perceived to be coming from roughly the same region of space, moving along similar spatial trajectories, and/or had common temporal synchrony and modulations in stimulus intensity or change. These correlated physical changes in photic and acoustic energy attributes are likely to serve to naturally bind audio-visual interactions, consistent with bottom-up Hebbian-like learning mechanisms. Such stimuli preferentially recruited circuitry of the bilateral pSTS complexes (Fig. 3E, blue vs pink), as addressed earlier.

In direct contrast to dynamic-visual stimuli, static-visual images (e.g. pictures, characters, and drawings) can have symbolic congruence with sound that must be learned to be associated with, and having few or no cross-modal invariant cues, thereby placing greater emphasis on declarative memory and related semantic-level matching mechanisms. The dynamic versus static visual stimulus dimension was further assessed using a subset of natural-only versus artificial stimuli. While there were insufficient numbers of studies in three of the sub-groups for definitive meta-analysis results (data not shown), the outcomes suggested a bias for the dynamic-visual stimuli clusters being driven by natural stimuli while the static-visual stimuli clusters may have been driven more by images involving relatively artificial stimuli (e.g. checkerboards, dots, circles, texture patterns). Regardless, a double-dissociation was evident.

Another consideration regarding the dynamic/natural versus static/artificial processing was depth-of-encoding. The greater depth required for encoding for subordinate versus basic level information is reported to recruit greater expanses of cortices along the anterior temporal lobes (Adams & Janata, 2002; Tranel et al., 2003; Tyler et al., 2004). For instance, associating a picture of an iconic dog to the sound “woof” represents a *basic*

level of semantic matching, while matching the specific and more highly familiar image of one's pet Tibetan terrier to her particular bark to be let outside would represent a *subordinate* level of matching that is regarded as having greater depth in its encoding. Neuroimaging and neuropsychological studies of semantically congruent cross-modal processing has led to a Conceptual Structure Account model (Taylor et al., 2009; Tyler & Moss, 2001) suggesting that objects in different categories can be characterized by the number and statistical properties of their constituent features (i.e. its depth), and this model points to the anterior temporal poles as “master binders” of such audio-visual information.

Correlating static-visual images with sound could be argued to require a more cognitive learning process than perceptually observing dynamic-visual events as they unfold and provide more intermodal-invariant information correlated with ongoing acoustic information. Thus, it was somewhat surprisingly that the static-visual stimuli preferentially recruited of the bilateral planum temporale (Fig. 3E, pink hues), in locations close to secondary auditory cortices, rather than in the temporal poles. However, this may relate to depth-of-encoding issues. The audio-visual stimuli used in many of the included used a relatively basic level of semantic matching (stimuli and tasks), which may have masked more subtle or widespread activation in inferotemporal cortices (e.g. temporal poles).

One possibility is that the pSTS complexes may represent intermediate processing stages that convey dynamically matched audio-visual congruent interaction information to the temporal poles, while the bilateral planum temporale regions may represent parallel intermediate processing stages that convey semantically congruent audio-visual

information derived from learned associations of sound with static (iconic) images referring to their matching source. Overall, this interpretation supports the tenet from unisensory systems “*that parallel hierarchical pathways process increasing information content*”, but here including two parallel multisensory processing pathways mediating the perception of audio-visual interaction information as events that are physically matched from a bottom-up perspective versus learned to be semantically congruent.

C. Semantic processing and neurocomputational models of cognition. Several mechanistic models regarding how and why semantic knowledge formation might develop in the brain includes the concept of hubs (and connector hubs) in brain networks (A. R. Damasio, 1999; Pulvermuller, 2018; Sporns et al., 2007), which are thought to allow for generalizations and the formation of categories. As such, the roughly six basic ROIs emerging from the present meta-analysis study (left and right pSTS complexes, left and right planum temporale, left fusiform, and right anterior insula) were of particular interest.

With regard to double-dissociations of cortical function, the right anterior insula and left fusiform ROIs had relatively small volumes, and thus may be considered less robust by some meta-analysis standards (also see Limitations below). Nonetheless, these preliminary findings provide at least moderate support for a taxonomic neurobiological model for processing different categories of real-world audio-visual events, which is readily amenable to testing with neurocomputational models and future hypothesis-driven multisensory processing studies. For instance, one might directly assess whether the different ROIs have functionally distinct characteristics as connector hubs for semantic

processing with activity dynamics that are functionally linking action perception circuits at a category level (Pulvermuller, 2018). Additionally, one may test for functional connectivity pattern differences across these ROIs (e.g. resting state functional connectivity MRI) in neurotypical individuals relative to various clinical populations. Overall, the results indicating that different semantic categories of audio-visual interaction events may be differentially processed along different brain regions supports the tenet that “*categorical perception emerges in neurotypical listeners [observers]*”, but here applying to the realm of cortical representations mediating multisensory object information. It remains unclear, however, whether this interpretation regarding categorical perception would provide greater support for domain-specific theoretical models, as proposed for some vision-dominated categories, such as the processing of faces, tools, fruits and vegetables, animals, and body parts (Caramazza & Shelton, 1998; H. Damasio et al., 1996; Mahon, Anzellotti, Schwarzbach, Zampini, & Caramazza, 2009; Mahon & Caramazza, 2005; Pascual-Leone & Hamilton, 2001) or for sensory-motor property-based models that develop because of experience (Barsalou, 2008; Barsalou, Kyle Simmons, Barbey, & Wilson, 2003; Lissauer, 1890/1988; Martin, 2007), or perhaps some combination of both.

D. Limitations. While this meta-analysis study revealed significant dissociations of cortical regions involved in different aspects of audio-visual interaction processing, at a more detailed or refined level there were several limitations to consider. As with most meta-analyses, the reported results were confined only to published ‘positive’ results, and tended to be biased in examining topics (in this case sensory stimulus categories) that

typically have greater rationale for being studied (and funded). In particular, the categories of living things (humans) and/or vocalizations (speech) are simply more thoroughly studied as socially- and health-relevant topics relative to the categories of non-living and non-vocal audio-visual stimuli, as evident in the numbers of studies listed in the provided tables. Because there were fewer numbers of studies in some semantic categories, double-dissociation differences could only be observed in some contrast meta-analyses when using uncorrected p-values, a statistical correction process that to date remains somewhat contentious in the field of meta-analyses. The biases in stimuli commonly used also led to the limitation that there would be greater heterogeneity of, for instance, non-living audio-visual sources and action events devoid of any vocalizations. This precluded examination of sub-categories such as environmental sources, mechanical (human-made) audio-visual sources, versus ‘artificial’ events (being computer-derived or other illusory sources), which limited a more thorough testing of the taxonomic model (Fig. 1) being investigated.

At a more technical level, other potential limitations included methodological differences across study designs, such as (a) differences in alignment methods, (b) imaging large swaths of brain rather than truly ‘whole brain’ imaging, and (c) potential inclusion of participants in more than one published study (which was not accessible information). Together, these limitations may constitute violations of assumptions by the ALE meta-analysis processes. Nonetheless, the modest support for our first two hypotheses and strong support for our third hypothesis should merit future study to validate and/or refine these basic cortical organization tenets and neurobiological taxonomic model.

Conclusions

This study summarized evidence derived from meta-analyses across 137 experimental paradigms to test for brain organizations for representing putative taxonomic boundaries related audio-visual perception of multisensory events at a category-level. The semantic categories tested were derived from an ethologically and evolutionarily inspired neurobiological model of real-world auditory event perception. The outcomes provided novel, though tentative support for the existence of double-dissociations mediating processing and perception around semantic categories, including (1) living versus non-living (artificial) audio-visual events, and (2) vocalization versus action audio-visual events. The outcomes further provided strong support for a double-dissociation for processing (3) dynamic-visual (mostly natural events) versus static-visual (including artificial) audio-visual interactions. Together, these findings were suggestive of parallel hierarchical pathways for processing and representing different semantic categories of multisensory event types, with embodiment strategies as potential underlying neuronal mechanisms. Overall, the present findings highlighted where and how auditory and visual perceptual representations interact in the brain, including the identification of a handful of cortical hubs in Figure 3C-E that are amenable to future neurocomputational modeling and testing of semantic knowledge representation mechanisms. Exploration of these and other potential multisensory hubs will be important for future studies addressing *why* specific brain regions may typically develop to process different aspects of audio-visual information, and thereby establish and maintain the “multisensory brain”, which

ultimately subserves many of the complexities of human communication and social behavior.

Availability of Data

The data that support the findings of this study are available from the corresponding author upon reasonable request.

UNCORRECTED MANUSCRIPT

FIGURES and TABLES

	LIVING THINGS		NON-LIVING THINGS	
	(conspecifics)	(non-conspecifics)		
ACTION SOUNDS (non-vocalizations)	Tool use sounds / Human action sounds	Animal action sounds	Environmental sounds	Mechanical sounds (human made)
VOCALIZATIONS	Human speech / Human non-verbal vocals	Animal vocal sounds		

Figure 1. A neurobiological model of the organization of the human brain for processing and recognizing different acoustic-semantic categories of natural sounds (from Breczynski-Lewis and Lewis, 2017). Bold text in the boxed regions depict rudimentary sound categories, including living versus non-living things and vocalizations versus non-vocal action sounds, which are categories being tested in the present audio-visual meta-analyses. Other sub-categories are also indicated, including human speech, tool use sounds, and human-made machinery sounds. Vocal and instrumental music sounds/events are regarded as higher forms of communication, which rely on other networks and are thus outside the scope of the present study. Refer to text for other details.

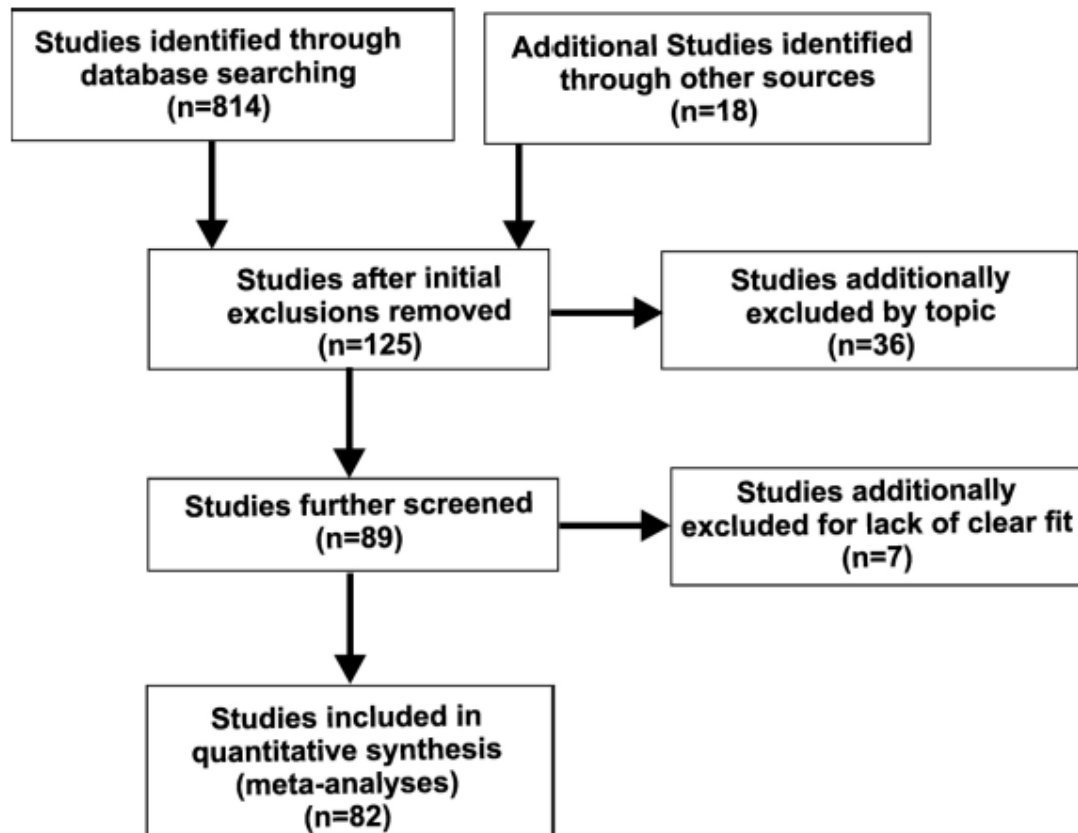


Figure 2. PRISMA table illustrating the flow of information through the different phases of the meta-analysis review.

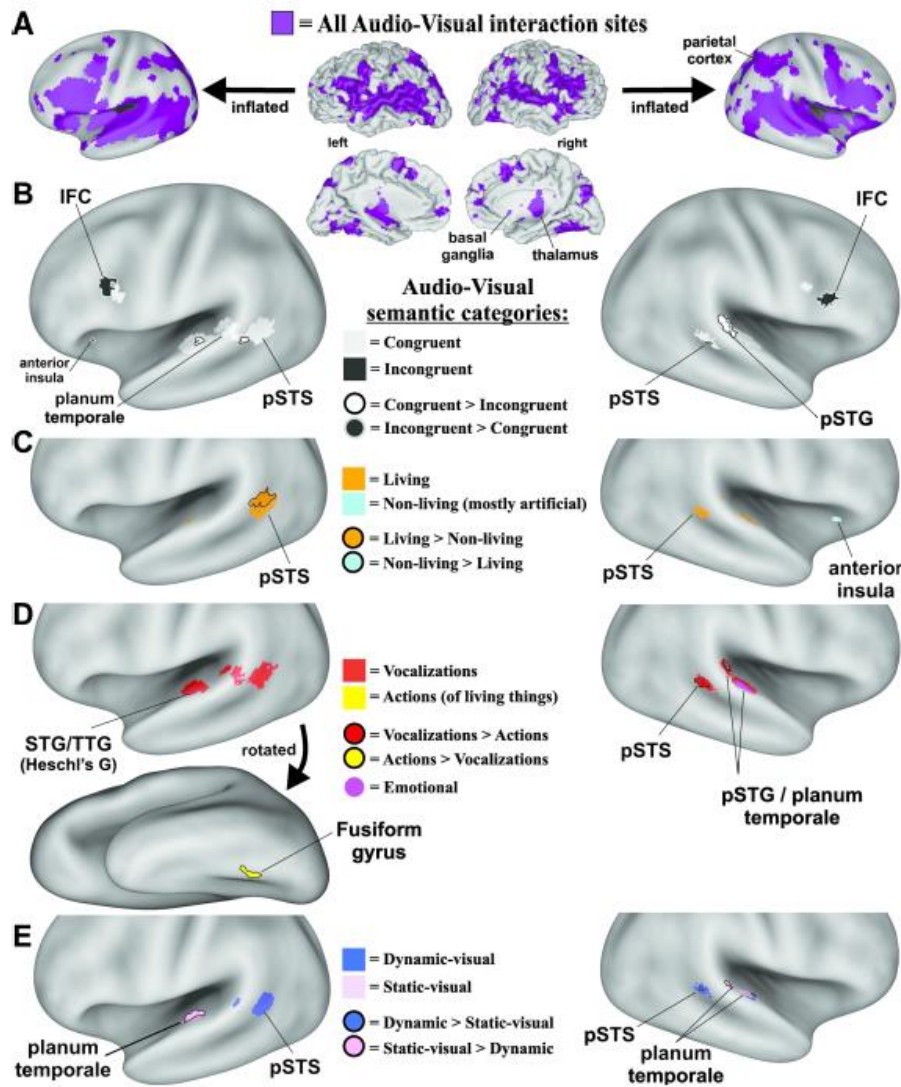


Figure 3. Activation likelihood estimate (ALE) maps of audio-visual interaction

sites. (A) Cortical maps derived from all studies (Table 1; purple hues, unthresholded) to illustrate global expanses of cortices involved. Data were projected onto the fiducial (lateral and medial views) and inflated (lateral views only) PALS atlas model of cortex.

(B) Illustration of maps derived from single study Congruent paradigms (yellow) plus superimposed maps of single study Incongruent audio-visual paradigms (green). Outlined foci depict ROIs after direct contrasts (e.g. Congruent > Incongruent). (C) ALE map

revealing single study Living (orange) contrasted with single study Non-living (cyan) categories of audio-visual paradigms. **(D)** ALE maps revealing audio-visual interactions involving single study Vocalizations (red) versus single study Action (mostly non-vocal) living source sounds (yellow). A single study ALE map for paradigms using Emotionally valent audio-visual stimuli, predominantly involving human vocalizations, is also illustrated (violet). **(E)** ALE maps showing ROIs preferentially recruited using single study dynamic-visual (blue hues) relative to single study static-visual (pink hues) audio-visual interaction foci. All single study ALE maps were at corrected FWE $p < .05$, and subsequently derived contrast maps were at uncorrected $p < .05$. IFC=inferior frontal cortex, pSTS = posterior superior temporal sulcus. TTG = transverse temporal gyrus. Refer to text for other details.

<Table 1. Paradigms cited. Note to **Reviewers** and **typesetter(s)**; the listing below of the included n=82 citations are shown here for reference tracking purposes in Tables using EndNote software and can be deleted for typesetting>. See table legend below.

(Adams & Janata, 2002) (Alink, Singer, & Muckli, 2008) (Balk et al., 2010) (Baumann & Greenlee, 2007) (Baumgaertner, Buccino, Lange, McNamara, & Binkofski, 2007) (Beauchamp, Argall, et al., 2004) (Beauchamp, Lee, et al., 2004) (Belardinelli et al., 2004) (Biau, Moris Fernandez, Holle, Avila, & Soto-Faraco, 2016) (Bischoff et al., 2007) (Blank & von Kriegstein, 2013) (Bonath et al., 2013) (Bonath et al., 2014) (Bushara, Grafman, & Hallett, 2001) (Bushara et al., 2003) (Callan, Jones, & Callan, 2014) (Calvert et al., 1999) (Calvert et al., 2000) (Calvert, Hansen, Iversen, & Brammer, 2001) (Calvert & Campbell, 2003) (de Haas, Schwarzkopf, Urner, & Rees, 2013) (Erickson et al., 2014) (Ethofer et al., 2013) (Gonzalo, Shallice, & Dolan, 2000) (Hagan, Woods, Johnson, Green, & Young, 2013) (Hasegawa et al., 2004) (Hashimoto & Sakai, 2004) (He et al., 2015) (Hein et al., 2007) (Hocking & Price, 2008) (Hove, Fairhurst, Kotz, & Keller, 2013) (James & Gauthier, 2003) (James, VanDerKlok, Stevenson, & James, 2011) (Jessen & Kotz, 2015) (Jola et al., 2013) (Kim et al., 2015) (Kreifelts, Ethofer, Grodd, Erb, & Wildgruber, 2007) (Lewis, Beauchamp, & DeYoe, 2000) (Matchin, Groulx, & Hickok, 2014) (McNamara et al., 2008) (Meyer, Baumann, Marchina, & Jancke, 2007) (Muller, Cieslik, Turetsky, & Eickhoff, 2012) (Murase et al., 2008) (Naghavi, Eriksson, Larsson, & Nyberg, 2011) (Nath & Beauchamp, 2012) (Naumer et al., 2008) (Naumer et al., 2011) (Noppeney et al., 2008) (Ogawa & Macaluso, 2013) (Okada, Venezia, Matchin, Saberi, & Hickok, 2013) (Olson, Gatenby, & Gore, 2002) (Plank et al., 2012) (Raij, Uutela, & Hari, 2000) (Robins, Hunyadi, & Schultz, 2009) (Scheef et al., 2009) (Schmid, Buchel, & Rose, 2011) (Sekiyama, Kanno, Miura, & Sugita, 2003) (Sestieri et al., 2006) (Stevenson & James, 2009) (Tanabe, Honda, & Sadato, 2005) (Taylor et al., 2006) (N. van Atteveldt, Formisano, Goebel, & Blomert, 2004) (N. M. van Atteveldt, Formisano, Goebel, & Blomert, 2007) (N. M. van Atteveldt, Blau, Blomert, & Goebel, 2010) (Vander Wyk et al., 2010) (von Kriegstein & Giraud, 2006) (Watkins, Shams, Tanaka, Haynes, & Rees, 2006) (Watkins, Shams, Josephs, & Rees, 2007) (Watson, Latinus, Noguchi, et al., 2014) (Watson, Latinus, Charest, Crabbe, & Belin, 2014) (Werner & Noppeney, 2010) (Wolf et al., 2014) (Green et al., 2009) (He et al., 2018) (Kircher et al., 2009) (Ogawa, Bordier, & Macaluso, 2013) (Regenbogen et al., 2018) (Straube, Green, Bromberger, & Kircher, 2011) (Straube et al., 2014) (Szyzik, Jansma, & Munte, 2009) (Willems, Ozyurek, & Hagoort, 2007)

Table 1. List of all studies used in the subsequent subsets of audio-visual interaction

site meta-analyses. The first column denotes the 82 included studies, and the second column the 137 experimental paradigms of those studies. The next columns depict first author (alphabetically), the year, and abbreviated description of the data table (T) or figure (F) used, followed by number of subjects. The column labeled “Multiple experiments” indicates that the multiple experimental paradigms where subject numbers were pooled from that study for the meta-analysis, such as for the single study ALE meta-analysis depicted in Fig. 3A (purple). The number of reported foci in the left and right hemispheres and their sum is also indicated. This is followed by a brief description of the experimental paradigm: B/W = black and white, A = audio, AV = audio-visual, V = visual, VA = visual-audio. The rightmost columns show the coding of experimental paradigms that appear in subsequent meta-analyses and Tables, with correspondence to the results illustrated in Figure 3: 0 = not used in contrast, 1= included, 2 = included as the contrast condition, blank cell = uncertain of clear category membership and not used in that contrast condition. See text for other details.

Study #	Experiment #	First author	Year	Experimental code and abbreviated task	# Subjects	Multiple experiments	Left hem foci	Right hem foci	Number of foci	Brief description of experimental paradigm	Congruent vs Incongruent	Living vs Non-living	Emotional stimuli/task	Vocalizations vs Not	Dynamic (video) vs static
82	137				1285		376	338	714		2B	2C	2C	2D	2E
1	1	Adams	2002	Expt 1 Table 3 A+V (aud coords only)	12		5	1	6	A and V commonly showing subordinate > basic object name verification (words with pictures or environmental sounds)	1	2		2	2
2	2	Alink	2008	Table 1c spheres move to drum sounds	10		4	6	10	Visual spheres and drum sounds moving: crossmodal dynamic capture vs conflicting motion	1	2		2	1
3	3	Balk	2010	Fig 2 asynchronous vs simultaneous	14		2	1	3	Natural asynchronous vs simultaneous AV speech synchrony (included both contrasts as interaction effects)	1			1	1
4	4	Baumann	2007	Table 1B coherent V+A vs A	12		2	1	3	Visual dots 16% coherent motion & in-phase acoustic noise > stationary acoustic sound	1	2	2	2	1
5	5	Baumann	2007	Table 2B		pooled	15	12	27	Moving acoustic noise & visual dots 16% in-phase coherent > random dot motion	1	2	2	2	1
6	6	Baumgaertner	2007	Table 3 Action > non-act sentence+video	19		3	0	3	Conjunction spoken sentences (actions-non-actions) AND videos (actions-non-actions)	1	1		1	1
7	7	Beauchamp	2004a	Fig 3J-K. Table 1 first 2 foci only	26		2	0	2	See photographs of tools, animals and hear corresponding sounds vs scrambled images, and synthesized rippled sounds	1	1	2	2	2
8	8	Beauchamp	2004b	Expt 1 coordinates	8		1	1	2	High resolution version of 2004a study: AV tool videos vs unimodal (AV > A,V)	1	1	2	2	1
9	9	Belardinelli	2004	Table 1 AV semantic congruence	13		6	6	12	Colored images of tools, animals, humans and semantically congruent vs incongruent sounds	1	1		1	2
10	10	Belardinelli	2004	Table 2 AV semantic incongruent		pooled	2	3	5	Colored images of tools, animals, humans and semantically incongruent vs congruent sounds	2	0	0	0	0
9	11	Biau	2016	Table 1A Interaction; speech synchronous	17		8	0	8	Hand gesture beats versus cartoon disc and speech interaction: synchronous vs asynchronous	1	1		1	1
10	12	Bischoff	2007	Table 2A only p<0.05 included	19		2	1	3	Ventriloquism effect: gray discs and tones, synchronous (p<0.05 corrected)	1			2	
11	13	Blank	2013	Fig 2	19		1	0	1	Visual-speech recognition correlated with recognition performance	1	1		1	1
12	14	Bonath	2013	pg 116 congruent thalamus	18		1	0	1	Small checkerboards and tones: spatially congruent vs incongruent (thalamus)	1	2	2	2	
15	15	Bonath	2013	pg 116 incongruent		pooled	1	1	2	Small checkerboards and tones: spatially incongruent vs congruent (thalamus)	2	0	0	0	0
13	16	Bonath	2014	Table 1A illusory vs not	20		1	5	6	Small checkerboards and tones: temporal > spatial congruence	1	2	2	2	
17	17	Bonath	2014	Table 1B synchronous > no illusion		pooled	3	0	3	Small checkerboards and tones: spatial > temporal congruence	1	2	2	2	
14	18	Bushara	2001	Table 1A (Fig 2) AV-Control	12		1	3	4	Tones (100 ms) & colored circles synchrony: detect Auditory then Visual presentation vs Control	1	2		2	
19	19	Bushara	2001	Table 1B (VA-C) five coords		pooled	2	3	5	Tones (100 ms) & colored circles synchrony: detect Visual then Auditory presentation vs Control	1	2		2	
20	20	Bushara	2001	Table 2A interact w/ Rt Insula		pooled	2	4	6	Tones & colored circles: correlated functional connections with (and including) the right insula	1	2		2	
15	21	Bushara	2003	Table 2A collide > pass, strong A-V interact	7		5	3	8	Tone and two visual bars moving: Tone synchrony induce perception they collide (AV interaction) vs pass by	1	2		2	1
16	22	Callan	2014	Table 5 AV-Audio (AV10-A10)-(AV6-A6)	16		4	4	8	Multisensory enhancement to visual speech in noise correlated with behavioral results	1	1		1	1
23	23	Callan	2014	Table 6 AV - Visual only		pooled	1	1	2	Multisensory enhancement to visual speech audio-visual versus visual only	1	1		1	1
17	24	Callan	1999	Table 1 (Fig 1)	5		3	4	7	View image of lower face and hear numbers 1 through 10 vs unimodal conditions (AV > Photos, Auditory)	1	1	2	1	2
18	25	Calvert	2000	Fig 2 superadditive+subadditive AVspeech	10		1	0	1	Speech & lower face: supra-additive plus sub-additive effects (AV-congruent > A,V > AV-incongruent)	1	1	2	1	1
26	26	Calvert	2000	Table 1 supraditive AVspeech		pooled	4	5	9	Speech & lower face: supra-additive AV enhancement	1	1	2	1	1
27	27	Calvert	2000	Table 2 incongruent subadditive AVspeech		pooled	3	3	6	Speech & lower face: sub-additive AV response to incongruent AV inputs	2	0	0	0	0
19	28	Calvert	2001	Table 2 superadditive & response depression	10		4	11	15	B/W visual checkerboard reversing & white noise bursts: Synchronous vs not: supraditive and response depression	1	2		2	
29	29	Calvert	2001	Table 3A superadditive only		pooled	6	4	10	B/W visual checkerboard reversing & white noise bursts: Synchronous vs not: supraditive only	1	2		2	
30	30	Calvert	2001	Table 3B response depression only		pooled	3	4	7	B/W visual checkerboard reversing & white noise bursts: Synchronous vs not: response depression only	1	2		2	
20	31	Calvert	2003	Table 2A (Fig 3 blue)	8		13	8	21	Speech and lower face: Moving dynamic speech (phonemes) versus stilled speech frames	1	1	2	1	1
21	32	DeHaas	2013	Table 1A AVcong - Visual	15		3	3	6	Video clips of natural scenes (animals, humans): AV congruent vs Visual	1	1			1
33	33	DeHaas	2013	Table 1B V- AV incongruent		pooled	2	0	2	Video clips of natural scenes (animals, humans): Visual vs AV incongruent	2	0	0	0	0
22	34	Erickson	2014	Table 1A Congruent AV speech	10		2	2	4	McGurk effect (phonemes): congruent AV speech: AV>A and AV>V	1	1	2	1	1
35	35	Erickson	2014	Table 1B McGurk speech		pooled	2	0	2	McGurk speech effect (phonemes)	1	1	2	1	1
23	36	Ethofer	2013	Table 1C emotion	23		1	2	3	Audiovisual emotional face-voice integration	1	1	1	1	1
24	37	Gonzalo	2000	Table 1 AV > AVincon music & Chinese ideograms	14		1	1	2	Learn novel Kanji characters & musical chords, activity increases over time for consistent AV pairings	1			2	2
38	38	Gonzalo	2000	Table 2 inconsistent AV		pooled	4	4	8	Learn novel Kanji characters & musical chords, activity increases over time to inconsistent pairings	2	0	0	0	0
39	39	Gonzalo	2000	Table 3 AV consistent vs Aud		pooled	1	1	2	Learn novel Kanji characters & musical chords, learn consistent (vs inconsistent) pairings versus auditory only	1			2	2
25	40	Green	2009	Table 1 incongruent > congruent gesture-speech	16		4	5	9	Incongruent vs congruent gesture-speech	2	1		1	1
41	41	Green	2009	Table 4A Congruent gesture-speech > gesture or speech		pooled	1	0	1	Congruent gesture-speech vs gesture with unfamiliar speech and with familiar speech	1	1		1	1
26	42	Hagan	2013	Table 1 AV emotion, novel over time	18		5	3	8	Affective audio-visual speech: congruent AV emotion vs A, V; unique ROIs over time (MEG)	1			1	1
43	43	Hagan	2013	Table 2 AV emotion incongruent		pooled	1	5	6	Affective audio-visual speech: incongruent AV emotion vs A, V; unique ROIs over time (MEG)	2	0	0	0	0
27	44	Hasegawa	2004	Table 1A (well trained piano) AV induced by V-only	26		12	6	18	Piano playing: well trained pianists, mapping hand movements to sequences of sound	1	1		2	1
28	45	Hashimoto	2004	Table 1G (Fig 4B, red) Learning Hangul letters to sounds	12		2	1	3	Unfamiliar Hangul letters & nonsense words, learn speech vs tone/noise pairings	1			2	1
29	46	He	2015	Table 3C AV speech foreign (left MTG focus)	20		1	0	1	Intrinsically meaningful gestures with German speech: Gesture-German > Gesture-Russian, German speech only	1	1		1	1
30	47	He	2018	Table 2 gestures & speech integration	20		1	0	1	Gesture-speech integration: Bimodal speech-gesture vs unimodal gesture with foreign speech and vs unimodal	1	1	0	1	1
31	48	Hein	2007	Fig 2A AV incongruent	18		0	2	2	Familiar animal images & incorrect (incongruent) vocalizations (dog; meow) vs correct pairs	2	0	0	0	0
49	49	Hein	2007	Fig 2B AV-artificial/non-living		pooled	0	1	1	B/W images of artificial objects ("fribbles") and animal vocalizations versus unimodal A, V	1				2
50	50	Hein	2007	Fig 2C pSTS, pSTG, mSTG AV-cong		pooled	0	3	3	Familiar animal images & correct vocalizations (dog; woof-woof)	1	1		1	2
51	51	Hein	2007	Fig 3A incongruent		pooled	4	0	4	AV familiar incongruent vs unfamiliar artificial (red foci 1,5,8,9)	2	0	0	0	0
52	52	Hein	2007	Fig 3B Foci 2, 3, 4 (blue) artificial/non-living		pooled	3	0	3	Visual "Fribbles" & backward/underwater distorted animal sounds, learn pairings (blue foci 2,3,4)	1			1	2
53	53	Hein	2007	Fig 3C congruent living (green)		pooled	3	0	3	Familiar congruent living vs artificial AV object features and animal sounds (green foci 7, 8, 10)	1			1	2
32	54	Hocking	2008	pg 2444 verbal	18		2	0	2	(pSTS mask) Color photos, written names, auditory names, environmental sounds conceptually matched "amodal"	1	1			2
55	55	Hocking	2008	Table 3 incongruent simultaneous matching		pooled	8	10	18	Incongruent sequential AV pairs (e.g. see drum, hear bassdrum) vs congruent pairs	2	0	0	0	0
33	56	Hove	2013	pg 316 AV interaction putamen	14		0	1	1	Interaction between (beep > flash) vs (siren > moving bar); left putamen focus	1	2			

34	57	James	2003	Fig 2	12		0	1	1	Activation by visual objects ("Greebles") associated with auditory features (e.g. buzzes, screeches); (STG)	1				2
35	58	James	2011	Table 1A bi-modal (vs scrambled)	12		4	2	6	Video of human manual actions (e.g. sawing): Auditory and Visual intact versus scrambled, AV event selectivity	1	1	2	2	1
36	59	Jessen	2015	Table 1A emotion > neutral AV enhanced	17		1	1	2	Emotional multisensory whole body and voice expressions: AV emotion (anger and fear) > neutral expressions	1	1	1	1	1
	60	Jessen	2015	Table 1D fear > neutral AV enhanced		pooled	2	1	3	Emotional multisensory whole body and voice expressions: AV fear > neutral expressions	1	1	1	1	1
37	61	Jola	2013	Table 1C AVcondition dance	12		3	3	6	Viewing unfamiliar dance performance (tells a story by gesture) with vs without music; using intersubject correlation	1	1	1	2	1
38	62	Kim	2015	Table 2A AV>C speech semantic match	15		2	0	2	Moving audio-visual speech perception vs white noise and unopened mouth movements	1	1			1
39	63	Kircher	2009	Fig 3B gesture related activation increase	14		3	1	4	Bimodal gesture-speech vs gesture and vs speech	1	1			1
40	64	Kreifelts	2007	Table 1 voice-face emotion	24		1	2	3	Facial expression & intonated spoken words, judge emotion expressed (AV > A.V; p<0.05 only)	1	1	1	1	1
	65	Kreifelts	2007	Table 5 AV increase effective connectivity		pooled	2	4	6	Increased effectiveness connectivity with pSTS and thalamus during AV integration of non-verbal emotional information	1	1	1	1	1
41	66	Lewis	2000	Table 1	7		2	3	5	Compare speed of tone sweeps to visual dot coherent motion; Bimodal vs unimodal	1	2			2
42	67	Matchin	2014	Table 1 AV > Aud only (McGurk)	20		2	7	9	McGurk audio-visual speech: AV > A only	1	1	2	1	1
	68	Matchin	2014	Table 2 AV > Video only		pooled	9	6	15	McGurk audio-visual speech: AV > V only	1	1	2	1	1
	69	Matchin	2014	Table 3 MM > AV McGurk		pooled	7	4	11	McGurk Mismatch > AV speech integration	2	0	0	0	0
43	70	McNamara	2008	Table (BA44 and IPL)	12		2	2	4	Videos of meaningless hand gestures & synthetic tone sounds: Increases in functional connectivity with learning	1				2
44	71	Meyer	2007	Table 3 paired A+V vs null	16		3	2	3	Paired screen red flashes with phone ring: paired V (conditioned stimulus) and A (unconditioned) vs null events	1	2			2
	72	Meyer	2007	Table 4 CS+, learned AV association with V-only		pooled	4	6	10	Paired screen flashes with phone ring: View flashes after post-conditioned vs null events	1	2			2
45	73	Muller	2012	Table S1 effective connectivity changes	27		4	3	7	Emotional facial expression (groaning, laughing) AV integration and gating of information	1	1	1	1	2
46	74	Murase	2008	Fig 4 discordant > concordant AVinteraction	28		1	0	1	Audiodivisional speech (syllables) showing activity to discordant versus concordant stimuli: left-right-STG	2	1	2	1	1
47	75	Naghavi	2007	Fig 1C	23		0	3	3	B/W pictures (animals, tools, instruments, vehicles) & their sounds: Congruent vs Incongruent	1	3			3
48	76	Naghavi	2011	Fig 2A cong = incon	30		1	0	1	B/W drawings of objects (living and non) and natural sounds (barking, piano): congruent = incongruent encoding	0				2
	77	Naghavi	2011	Fig 2B congruent > incongruent		pooled	0	1	1	B/W drawings of objects (living and non) and natural sounds (barking, piano): congruent > incongruent encoding	1				2
	78	Naghavi	2011	Fig 2C incongruent > congruent		pooled	1	1	2	B/W drawings of objects (living and non) and natural sounds (barking, piano): incongruent > congruent encoding	2	0	0	0	0
49	79	Nath	2012	pg 784	14		1	0	1	McGurk effect (phonemes): congruent AV speech correlated with behavioral percept	1	1	2	1	1
50	80	Naumer	2008	Fig 2 Table 1A max contrast	18		8	6	14	Images of "Fribbles" & learned artificial sounds (underwater animal vocals): post training vs max contrast	1				2
	81	Naumer	2008	Fig 3 Table 1B pre-post		pooled	5	6	11	Images of "Fribbles" & learned corresponding artificial sounds: Post vs Pre-training session	1				2
	82	Naumer	2008	Fig 4 Table 2		pooled	1	1	2	Learn of "Fribbles" and distorted sounds as incongruent > congruent pairs	2	0	0	0	0
51	83	Naumer	2011	Fig 3C	10		1	0	1	Photographs of objects (living and non) and related natural sounds	1				2
52	84	Noppeny	2008	Table 2 AV incongruent > congruent	17		5	2	7	Speech sound recognition through AV priming, environmental sounds and spoken words: Incongruent > congruent	2	0	0	0	0
	85	Noppeny	2008	Table 3 AV congruent sounds/words		pooled	4	0	4	Speech sound recognition through AV priming, environmental sounds and spoken words: Congruent > incongruent	1				2
53	86	Ogawa	2013a	Table 1 (pg 162 data)	13		1	0	1	AV congruency of pure tone and white dots moving on screen (area left V3A)	1	2			2
54	87	Ogawa	2013b	Table 1 3D > 2D and surround > monaural effects	16		3	4	7	Cinematic 3D > 2D video and surround sound > monaural while watching a movie ("The Three Musketeers")	1	1			0
55	88	Okada	2013	Table 1 AV > A	20		5	4	9	Video of AV > A speech only	1	1			1
56	89	Olson	2002	Table 1A synchronized AV > static Vis-only	10		7	4	11	Whole face video & heard words: Synchronized AV vs static V	1	1			1
	90	Olson	2002	Table 1C, synchronized AV > desynchronized AV speech		pooled	2	0	2	Whole face video & heard words: Synchronized vs desynchronized	1	1			1
57	91	Plank	2012	pg 803 AV congruent effect	15		0	1	1	AV spatially congruent > semantically matching images of natural objects and associated sounds (right STG)	1	3			3
	92	Plank	2012	Table 2A spatially congruent-baseline		pooled	5	5	10	Images of natural objects and associated sounds, spatially congruent vs baseline	1	3			3
58	93	Rajj	2000	Table 1B letters & speech sounds	9		2	3	5	Integration of visual letters and corresponding auditory phonetic expressions (MEG study) AV vs (A + V)	1		2	1	2
59	94	Regehbogen	2017	Table 2A degraded > clear Multisensory vs unimodal input	29		5	6	11	Degraded > clear AV vs both visual and auditory unimodal visual real-world object-in-action recognition	1	0			2
60	95	Robins	2008	Table 2 (Fig 2) AV integration (AV>A and AV>V)	10		2	1	3	Face speaking sentences: angry, fearful, happy, neutral (AV > A, V)	1				1
	96	Robins	2008	Table 4A (Fig 5) AV integration & emotion		pooled	1	4	5	AV faces and spoken sentences expressing fear or neutral valence: AV integration (AV > A, V conditions)	1	1			1
	97	Robins	2008	Table 4B emotion effects		pooled	2	0	2	AV faces and spoken sentences expressing fear or neutral valence: Emotional AV-fear > AV-neutral	1	1	1	1	1
	98	Robins	2008	Table 4C (Fig 5) fearful AV integration		pooled	1	5	6	AV faces and spoken sentences expressing fear or neutral valence: Fearful-only AV integration	1	1	1	1	1
	99	Robins	2008	Table 4D AV-only emotion		pooled	1	3	4	AV faces and spoken sentences expressing fear or neutral valence: AV-only emotion	1	1	1	1	1
61	100	Scheef	2009	Table 1 cartoon jump + boing	16		1	2	3	Video of cartoon person jumping & "sonification" of a tone, learn correlated pairings: AV-V and AV-A conjunction	1	1	2	2	1
62	101	Schmid	2011	Table 2E A effect V (Living & non-living, pictures)	12		3	4	7	Environmental sounds & matching pictures: reduced activity by A	1	3			3
	102	Schmid	2011	Table 2F V competition effect A (reduced activity by a visual object)		pooled	2	2	4	Environmental sounds & matching pictures: reduced activity by V	1	3			3
	103	Schmid	2011	Table 2G AV crossmodal interaction x auditory attention		pooled	2	3	5	Environmental sounds and matching pictures: cross-modal interaction and auditory attention	1	3			3
63	104	Sekiyama	2003	Table 3 (fMRI nAV-AV)	8		1	0	1	AV speech, McGurk effect with phonemes (ba, da, ga) and noise modulation: noise-AV > AV (fMRI)	1	1	2	1	1
	105	Sekiyama	2003	Table 4 (PET nAV-AV)		pooled	1	3	4	AV speech, McGurk effect with phonemes (ba, da, ga) and noise modulation: noise-AV > AV (PET)	1	1	2	1	1
64	106	Sestieri	2006	Table 1 (Fig 3), AV location match vs semantic	10		2	5	7	B/W images (animal, weapons) & environmental sounds: Match location > recognition	1	1			2
	107	Sestieri	2006	Table 2 AV semantic recognition vs localization		pooled	2	1	3	B/W pictures and environmental sounds: congruent semantic recognition > localization task	1	3			3
65	108	Stevenson	2009	Table 1B AVtools > AVspeech	11		1	1	2	Hand tools in use video: inverse effectiveness (degraded AV tool > AV speech)	1	1	2	3	1
	109	Stevenson	2009	Table 1C (Fig 8) AVspeech > AVtools		pooled	1	1	2	Face & speech video: inverse effectiveness (degraded AV speech > AV tool use)	1	1			1
66	110	Straube	2011	Table 3A-B iconic/metaphoric speech-gestures vs speech, gestu	16		2	2	4	Integration of Iconic and Metaphoric speech-gestures vs speech and gesture	1	1			1
67	111	Straube	2014	pg39 Integration foci	16		3	0	3	Integration of iconic hand gesture-speech > unimodal speech and unimodal gesture (healthy control group)	1	1			1
68	112	Szyck	2009	Table 1 AV incongruent > AV congruent face+speech	11		7	2	9	Incongruent AV face-speech vs congruent AV face-speech	2	1			1

69	113	Tanabe	2005	Table 1A. AV; A then V; not VA	15		10	10	20	Amorphous texture patterns & modulated white noises: Activation during learning delay period (AV)	1	2	2	2	2
	114	Tanabe	2005	Table 2A+2B (Fig 5a) AV and VA		pooled	5	6	11	Amorphous texture patterns & modulated white noises: changes after feedback learning (AV and VA)	1	2	2	2	2
	115	Tanabe	2005	Table 3A+3B (Fig 6) AV and VA; delay period		pooled	9	1	10	Amorphous texture patterns & modulated white noises: sustained activity throughout learning (AV and VA)	1	2	2	2	2
70	116	Taylor	2006	pg 8240 AV incongruent	15		1	0	1	Color photos (V), environmental sounds (A), spoken words: Incongruent (living objects)	2	0	0	0	0
	117	Taylor	2006	Fig 1A-B, Fig 1C-D (living > non-living)		pooled	2	0	2	Color photos (V), environmental sounds and spoken words (A): Cong AV vs Incong (living objects)	1				2
71	118	Van Atteveldt	2004	Table 1a letters & speech sounds	16		3	1	4	Familiar letters & their speech sounds: Congruent vs not and Bimodal vs Unimodal	1			1	2
72	119	Van Atteveldt	2007	Table 2A+B (Fig 2)	12		3	2	5	Single letters & their speech sounds (phonemes): Congruent > Incong; Passive perception, blocked and event-related	1			1	2
	120	Van Atteveldt	2007	Table 3 (Fig 2) passive		pooled	1	1	2	Single letters & their speech sounds (phonemes): Congruent > Incong; active perception task	1			1	2
	121	Van Atteveldt	2007	Table 4 (Fig 6) active condition, incongruent		pooled	1	6	7	Single letters & their speech sounds (phonemes): Incongruent > Congruent	2	0	0	0	0
73	122	Van Atteveldt	2010	Table 1B STS: specific adaptation congruent > incong	16		3	1	4	Letter and speech sound pairs (vowels, consonants): Specific adaptation effects	1			1	2
74	123	Van der Wyk	2010	Table 2 AV interaction effects oval/circles+speech/nonspeech	16		3	3	6	Geometric shape modulate with speech (sentences)	1			1	1
75	124	Von Kriegstein	2006	Fig 4B after > before voice-face	14		0	4	4	Face and object photos with voice and other sounds: Voice-Face association learning	1	1		1	2
76	125	Watkins	2006	Fig 4 illusory multisensory interaction	11		0	2	2	Two brief tone pips leads to illusion of two screen flashes (annulus with checkerboard) when only one flash present	1	2		2	
	126	Watkins	2006	Table 1 (A enhances V in general)		pooled	5	3	8	Single brief tone pip leads to illusion of single screen flash (annulus with checkerboard) when two flashes present	1	2		2	
77	127	Watkins	2007	Fig 3 2 flashes + 1 beep illusion	10		0	1	1	Two visual flashes and single audio beep leads to the illusion of a single flash	1	2		2	
78	128	Watson	2014a	Table 1A. AV-adaptation effect (multimodal localizer)	18		0	1	1	Videos of emotional faces and voice; multisensory localizer	1	1		1	1
	129	Watson	2014a	Table 1C. AV-adaptation effect, cross-modal adaptation effect		pooled	0	1	1	Videos of emotional faces and voice; cross-modal adaptation effects	1	1	1	1	1
79	130	Watson	2014b	Table 1 AV > baseline (Living and non-Living)	40		3	5	8	Moving objects and videos of faces with corresponding sounds: AV > baseline	1				1
	131	Watson	2014b	Table 4A integrative regions (Living and non-Living)		pooled	2	2	4	Moving objects and videos of faces with corresponding sounds: Integrative regions (AV > A,V)	1				1
	132	Watson	2014b	Table 4B integrative regions (Living and non-Living)		pooled	0	1	1	Moving objects and videos of faces with corresponding sounds: People selective integrative region	1	1		1	1
80	133	Werner	2010	Table 1 superadditive (AV-salience effect)	21		0	3	3	Categorize movies of actions with tools or musical instruments (degraded stimuli); AV interactions both tasks	1	1	2	2	1
	134	Werner	2010	Table 2 AV interactions predict behavior		pooled	1	2	3	Categorize movies of actions with tools or musical instruments; AV interactions predicted by behavior	1	1	2	2	1
	135	Werner	2010	Table 3C superadditive AV due to task		pooled	3	0	3	Categorize movies of actions with tools or musical instruments; Subadditive AV to task	1	1	2	2	1
81	136	Willems	2007	Table 3C-D mismatch hand gestures and speech	16		2	1	3	Mismatch of hand gesture (no face) and speech versus correct	2	1		1	1
82	137	Wolf	2014	Table 1 face cartoons + phonemes	16		1	1	2	Drawing of faces with emotional expressions: Supramodal effects with emotional valence	1	1	1	1	2

UNCORRECTED

Downloaded from https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/igab002/6103812 by West Virginia University user on 05 February 2021

Table 2. Studies included in the Congruent category for audio-visual interaction site

meta-analyses. This meta-analysis included 79 of the studies with 117 experiments (columns A and B). The column “multiple experiments” indicates paradigms where the same set of participants were included, and so all coordinates were pooled together as one study to avoid biases related to violation the assumption of subject independence (refer to Methods). Results of Congruent meta-analyses are shown in Fig. 3B white hues. Refer to Table 1 and text for other details.

UNCORRECTED MANUSCRIPT

Study #	Experiment #	First author	Year	Experimental code and abbreviated task	# Subjects	Multiple experiments	Left hemi foci	Right hemi foci	Number of foci	Brief description of experimental paradigm
79	117	Adams	2002	Expt 1 Table 3 A+V (aud coords only)	12		5	1	6	A and V commonly showing subordinate > basic object name verification (words with pictures or environmental sounds)
1	1	Alink	2008	Table 1c spheres move to drum sounds	10		4	6	10	Visual spheres and drum sounds moving: crossmodal dynamic capture vs conflicting motion
3	3	Balk	2010	Fig 2 asynchronous vs simultaneous	14		2	1	3	Natural asynchronous vs simultaneous AV speech synchrony (included both contrasts as interaction effects)
4	4	Baumann	2007	Table 1B coherent V+A vs A	12		2	1	3	Visual dots 16% coherent motion & in-phase acoustic noise > stationary acoustic sound
	5	Baumann	2007	Table 2B		pooled	15	12	27	Moving acoustic noise & visual dots 16% in-phase coherent > random dot motion
5	6	Baumgaertner	2007	Table 3 Action > non-act sentence+video	19		3	0	3	Conjunction spoken sentences (actions > non-actions) AND videos (actions > non-actions)
6	7	Beauchamp	2004a	Fig 3J-K, Table 1 first 2 foci only	26		2	0	2	See photographs of tools, animals and hear corresponding sounds vs scrambled images and synthesized rippled sounds
7	8	Beauchamp	2004b	Expt 1 coordinates	8		1	1	2	High resolution version of 2004a study: AV tool videos vs unimodal (AV > A,V)
8	9	Belardinelli	2004	Table 1 AV semantic congruence	13		6	6	12	Colored images of tools, animals, humans and semantically congruent vs incongruent sounds
9	11	Biau	2016	Table 1A Interaction: speech synchronous	17		8	0	8	Hand gesture beats versus cartoon disc and speech interaction: synchronous vs asynchronous
10	12	Bischoff	2007	Table 2A only p<0.05 included	19		2	1	3	Ventriloquism effect: gray discs and Jones, synchronous (p<0.05 corrected)
11	13	Blank	2013	Fig 2	19		1	0	1	Visual-speech recognition correlated with recognition performance
12	14	Bonath	2013	pg 116 congruent thalamus	18		1	0	1	Small checkerboards and tones: spatially congruent vs incongruent (thalamus)
13	16	Bonath	2014	Table 1A illusory vs not	20		1	5	6	Small checkerboards and tones: temporal > spatial congruence
	17	Bonath	2014	Table 1B synchronous > no illusion		pooled	3	0	3	Small checkerboards and tones: spatial > temporal congruence
14	18	Bushara	2001	Table 1A (Fig 2) AV-Control	12		1	3	4	Tones (100 ms) & colored circles synchrony: detect Auditory then Visual presentation vs Control
	19	Bushara	2001	Table 1B (VA-C) five coords		pooled	2	3	5	Tones (100 ms) & colored circles synchrony: detect Visual then Auditory presentation vs Control
	20	Bushara	2001	Table 2A interact w Rt Insula		pooled	2	4	6	Tones & colored circles: correlated functional connections with (and including) the right insula
15	21	Bushara	2003	Table 2A collide > pass, strong A-V interact	7		5	3	8	Tone and two visual bars moving: Tone synchrony induce perception they collide (AV interaction) vs pass by
16	22	Callan	2014	Table 5 AV-Audio (AV10-A10)-(AV6-A6)	16		4	4	8	Multisensory enhancement to visual speech in noise correlated with behavioral results
	23	Callan	2014	Table 6 AV - Visual only		pooled	1	1	2	Multisensory enhancement to visual speech audio-visual versus visual only
17	24	Calvert	1999	Table 1 (Fig 1)	5		3	4	7	View image of lower face and hear numbers 1 through 10 vs unimodal conditions (AV > Photos, Auditory)
18	25	Calvert	2000	Fig 2 superadd+subadd AV/speech	10		1	0	1	Speech & lower face: supra-additive plus sub-additive effects (AV-congruent > A,V > AV-incongruent)
	26	Calvert	2000	Table 1 supradd AV/speech		pooled	4	5	9	Speech & lower face: supra-additive AV enhancement
19	28	Calvert	2001	Table 2 superadditive & resp depression	10		4	11	15	B/W visual checkerboard reversing & white noise bursts: Synchronous vs not: supradditive and response depression
	29	Calvert	2001	Table 3A superadditive only		pooled	6	4	10	B/W visual checkerboard reversing & white noise bursts: Synchronous vs not: supradditive only
	30	Calvert	2001	Table 3B response depression only		pooled	3	4	7	B/W visual checkerboard reversing & white noise bursts: Synchronous vs not: response depression only
20	31	Calvert	2003	Table 2A (Fig 3 blue)	8		13	8	21	Speech and lower face: Moving dynamic speech (phonemes) versus skilled speech frames
21	32	DeHaas	2013	Table 1A AVcong - Visual	15		3	3	6	Video clips of natural scenes (animals, humans): AV congruent vs Visual
22	34	Erickson	2014	Table 1A Congruent AV speech	10		2	2	4	McGurk effect (phonemes): congruent AV speech: AV > A and AV > V
	35	Erickson	2014	Table 1B McGurk speech		pooled	2	0	2	McGurk speech effect (phonemes)
23	36	Ethofer	2013	Table 1C emotion	23		1	2	3	Audiovisual emotional face-voice integration
24	37	Gonzalo	2000	Table 1 AV > AVincon music & Chinese ideograms	14		1	1	2	Learn novel Kanji characters & musical chords, activity increases over time for consistent AV pairings
	39	Gonzalo	2000	Table 3 AV consistent vs Aud		pooled	1	1	2	Learn novel Kanji characters & musical chords, learn consistent (vs inconsistent) pairings versus auditory only
25	41	Green	2009	Table 4A Congruent gesture-speech > gesture or speech	16	pooled	1	0	1	Congruent gesture-speech vs gesture with unfamiliar speech and with familiar speech

Downloaded from https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/igab002/6103812 by West Virginia University user on 05 February 2021

26	42	Hagan	2013	Table 1 AV emotion, novel over time	18		5	3	8	Affective audio-visual speech: congruent AV emotion vs A, V; unique ROIs over time (MEG)
27	44	Hasegawa	2004	Table 1A (well trained piano) AV induced by V-only	26		12	6	18	Piano playing: well trained pianists, mapping hand movements to sequences of sound
28	45	Hashimoto	2004	Table 1G (Fig4B, red) Learning Hangul letters to sounds	12		2	1	3	Unfamiliar Hangul letters & nonsense words, learn speech vs tone/noise pairings
29	46	He	2015	Table 3C AV speech foreign (left MTG focus)	20		1	0	1	Intrinsically meaningful gestures with German speech: Gesture-German > Gesture-Russian, German speech only
30	47	He	2018	Table 2, GSI, left MTG, gestures & speech integration	20		1	0	1	Gesture-speech integration: Bimodal speech-gesture vs unimodal gesture with foreign speech and vs unimodal
31	49	Hein	2007	Fig 2B AV-artificial/non-living	18		0	1	1	B/W images of artificial objects ("fribbles") and animal vocalizations versus unimodal A, V
	50	Hein	2007	Fig 2C pSTS, pSTG, mSTG AV-cong		pooled	0	3	3	Familiar animal images & correct vocalizations (dog: woof-woof)
	52	Hein	2007	Fig 3B Foci 2, 3, 4 (blue) artificial/non-living		pooled	3	0	3	Visual "Fribbles" & backward/underwater distorted animal sounds, learn pairings (blue foci 2,3,4)
	53	Hein	2007	Fig 3C congruent living (green)		pooled	3	0	3	Familiar congruent living vs artificial AV object features and animal sounds (green foci 7, 8, 10)
32	54	Hocking	2008	pg 2444 verbal	18		2	0	2	(pSTS mask) Color photos, written names, auditory names, environmental sounds conceptually matched "amodal"
33	56	Hove	2013	pg 316 AV interaction putamen	14		0	1	1	Interaction between (beep > flash) vs (siren > moving bar), left putamen focus
34	57	James	2003	Fig 2	12		0	1	1	Activation by visual objects ("Greebles") associated with auditory features (e.g. buzzes, screeches), (STG)
35	58	James	2011	Table 1A bi-modal (vs scrambled)	12		4	2	6	Video of human manual actions (e.g. sawing), Auditory and Visual intact versus scrambled, AV event selectivity
36	59	Jessen	2015	Table 1A emotion > neutral AV enhanced	17		1	1	2	Emotional multisensory whole body and voice expressions: AV emotion (anger and fear) > neutral expressions
	60	Jessen	2015	Table 1D fear > neutral AV enhanced		pooled	2	1	3	Emotional multisensory whole body and voice expressions: AV fear > neutral expressions
37	61	Jola	2013	Table 1C AVcondition dance	12		3	3	6	Viewing unfamiliar dance performance (tells a story by gesture) with vs without music; using intersubject correlation
38	62	Kim	2015	Table 2A AV>C speech semantic match	15		2	0	2	Moving audio-visual speech perception vs white noise and unopened mouth movements
39	63	Kircher	2009	Figure 3B: gesture related activation increase	14		3	1	4	Bimodal gesture-speech vs gesture and vs speech
40	64	Kreifelts	2007	Table 1 voice-face emotion	24		1	2	3	Facial expression & intonated spoken words, judge emotion expressed (AV > A, V; p<0.05 only)
	65	Kreifelts	2007	Table 5 AV increase effective connectivity		pooled	2	4	6	Increased effectiveness connectivity with pSTS and thalamus during AV integration of non-verbal emotional information
41	66	Lewis	2000	Table 1	7		2	3	5	Compare speed of tone sweeps to visual dot coherent motion: Bimodal vs unimodal
42	67	Matchin	2014	Table 1 AV > Aud only (McGurk)	20		2	7	9	McGurk audio-visual speech: AV > A only
	68	Matchin	2014	Table 2 AV > Video only		pooled	9	6	15	McGurk audio-visual speech: AV > V only
43	70	McNamara	2008	Table (BA44 and IPL)	12		2	2	4	Videos of meaningless hand gestures & synthetic tone sounds: Increases in functional connectivity with learning
44	71	Meyer	2007	Table 3 paired A+V vs null	16		3	3	6	Paired screen red flashes with phone ring: paired V (conditioned stimulus) and A (unconditioned) vs null events
	72	Meyer	2007	Table 4 CS+, learned AV association with V-only		pooled	4	6	10	Paired screen flashes with phone ring: View flashes after post-conditioned vs null events
45	73	Muller	2012	Table S1 effective connectivity changes	27		4	3	7	Emotional facial expression (groaning, laughing) AV integration and gating of information
47	75	Naghavi	2007	Fig 1C	23		0	3	3	B/W pictures (animals, tools, instruments, vehicles) & their sounds: congruent vs incongruent
48	77	Naghavi	2011	Fig 2B congruent > incongruent	30		0	1	1	B/W drawings of objects (living and non) and natural sounds (barking, piano): congruent > incongruent encoding
49	79	Nath	2012	pg 784	14		1	0	1	McGurk effect (phonemes): congruent AV speech correlated with behavioral percept
50	80	Naumer	2008	Fig 2 Table 1A max contrast	18		8	6	14	Images of "Fribbles" & learned artificial sounds (underwater animal vocals); post training vs max contrast
	81	Naumer	2008	Fig 3 Table 1B pre-post		pooled	5	6	11	Images of "Fribbles" & learned corresponding artificial sounds: Post- vs Pre-training session
51	83	Naumer	2011	Fig 3C	10		1	0	1	Photographs of objects (living and non) and related natural sounds
52	85	Noppeny	2008	Table 3 AV congruent sounds/words	17		4	0	4	Speech sound recognition through AV priming, environmental sounds and spoken words: Congruent > incongruent
53	86	Ogawa	2013a	Table 1 (pg 162 data)	13		1	0	1	AV congruency of pure tone and white dots moving on screen (area left V3A)
54	87	Ogawa	2013b	Table 1 3D > 2D and surround > monaural effects	16		3	4	7	Cinematic 3D > 2D video and surround sound > monaural while watching a movie ("The Three Musketeers")
55	88	Okada	2013	Table 1 AV > A	20		5	4	9	Video of AV > A speech only
56	89	Olson	2002	Table 1A, synchronized AV > static Vis-only	10		7	4	11	Whole face video & heard words: Synchronized AV vs static V
	90	Olson	2002	Table 1C, synchronized AV > desynchronized AV speech		pooled	2	0	2	Whole face video & heard words: Synchronized vs desynchronized
57	91	Plank	2012	pg 803 AV congruent effect	15		0	1	1	AV spatially congruent > semantically matching images of natural objects and associated sounds (right STG)
	92	Plank	2012	Table 2A spatially congruent-baseline		pooled	5	5	10	Images of natural objects and associated sounds, spatially congruent vs baseline
58	93	Raij	2000	Table 1B letters & speech sounds	9		2	3	5	Integration of visual letters and corresponding auditory phonetic expressions (MEG study) AV vs (A + V)
59	94	Regenbogen	2017	Table 2A degraded > clear Multisensory vs unimodal input	29		5	6	11	Degraded > clear AV vs both visual and auditory unimodal visual real-world object-in-action recognition
60	95	Robins	2008	Table 2 (Fig 2) AV integration (AV>A and AV>V)	10		2	1	3	Face speaking sentences: angry, fearful, happy, neutral (AV > A, V)
	96	Robins	2008	Table 4A (Fig 5) AV integration & emotion	5		1	4	5	AV faces and spoken sentences expressing fear or neutral valence: AV integration (AV > A, V conditions)
	97	Robins	2008	Table 4B emotion effects		pooled	2	0	2	AV faces and spoken sentences expressing fear or neutral valence: Emotional AV-fear > AV-neutral
	98	Robins	2008	Table 4C (Fig 5) fearful AV integration		pooled	1	5	6	AV faces and spoken sentences expressing fear or neutral valence: Fearful-only AV integration
	99	Robins	2008	Table 4D AV-only emotion		pooled	1	3	4	AV faces and spoken sentences expressing fear or neutral valence: AV-only emotion

Downloaded from https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/igab002/6103812 by West Virginia University user on 05 February 2021

61	100	Scheef	2009	Table 1 cartoon jump + boing	16		1	2	3	Video of cartoon person jumping & "sonification" of a tone, learn correlated pairings: AV-V and AV-A conjunction
62	101	Schmid	2011	Table 2E A effect V (Living & non-living, pictures)	12		3	4	7	Environmental sounds & matching pictures: reduced activity by A
	102	Schmid	2011	Table 2F V competition effect A (reduced activity by a visual object)		pooled	2	2	4	Environmental sounds & matching pictures: reduced activity by V
	103	Schmid	2011	Table 2G AV crossmodal interaction x auditory attention		pooled	2	3	5	Environmental sounds and matching pictures: cross-modal interaction and auditory attention
63	104	Sekiyama	2003	Table 3 (fMRI nAV-AV)	8		1	0	1	AV speech, McGurk effect with phonemes (ba, da, ga) and noise modulation: noise-AV > AV (fMRI)
	105	Sekiyama	2003	Table 4 (PET nAV-AV)		pooled	1	3	4	AV speech, McGurk effect with phonemes (ba, da, ga) and noise modulation: noise-AV > AV (PET)
64	106	Sestieri	2006	Table 1 (Fig. 3), AV location match vs semantic	10		2	5	7	B/W images (animal, weapons) & environmental sounds: Match location > recognition
	107	Sestieri	2006	Table 2 AV semantic recognition vs localization		pooled	2	1	3	B/W pictures and environmental sounds: congruent semantic recognition > localization task
65	108	Stevenson	2009	Table 1B 2 AVtools > AVspeech	11		1	1	2	Hand tools in use video: inverse effectiveness (degraded AV tool > AV speech)
	109	Stevenson	2009	Table 1C (Fig. 8) AVspeech > AVtools		pooled	1	1	2	Face & speech video: inverse effectiveness (degraded AV speech > AV tool use)
66	110	Straube	2011	Table 3A-B iconic/metaphoric speech-gestures vs speech, gestu	16		2	2	4	Integration of iconic and Metaphoric speech-gestures vs speech and gesture
67	111	Straube	2014	p939 Integration foci	16		3	0	3	Integration of iconic hand gesture-speech > unimodal speech and unimodal gesture (healthy control group)
69	113	Tanabe	2005	Table 1A AV: A then V; not VA	15		10	10	20	Amorphous texture patterns & modulated white noises: Activation during learning delay period (AV)
	114	Tanabe	2005	Table 2A+2B (Fig 5a) AV and VA		pooled	5	6	11	Amorphous texture patterns & modulated white noises: changes after feedback learning (AV and VA)
	115	Tanabe	2005	Table 3A+3B (Fig. 6) AV and VA; delay period		pooled	9	1	10	Amorphous texture patterns & modulated white noises: sustained activity throughout learning (AV and VA)
70	117	Taylor	2006	Fig 1A-B, Fig 1C-D (living > non-living)	15		2	0	2	Color photos (V), environmental sounds and spoken words (A): Congruent AV vs incongruent (living objects)
71	118	Van Atteveldt	2004	Table 1a letters & speech sounds	16		3	1	4	Familiar letters & their speech sounds: Congruent vs not and Bimodal vs Unimodal
72	119	Van Atteveldt	2007	Table 2A+B (Fig 2)	12		3	2	5	Single letters & their speech sounds (phonemes): Congruent > Incong; Passive perception, blocked and event-related
	120	Van Atteveldt	2007	Table 3 (Fig 2) passive		pooled	1	1	2	Single letters & their speech sounds (phonemes): Congruent > Incong; active perception task
73	122	Van Atteveldt	2010	Table 1B STS: specific adaptation congruent > incong	16		3	1	4	Letter and speech sound pairs (vowels, consonants): Specific adaptation effects
74	123	Van der Wyk	2010	Table 2 AV interaction effects oval/circles+speech/nonspeech	16		3	3	6	Geometric shape modulate with speech (sentences)
75	124	Von Kriegstein	2006	Fig 4B after > before voice-face	14		0	4	4	Face and object photos with voice and other sounds: Voice-Face association learning
76	125	Watkins	2006	Fig 4 illusory multisensory interaction	11		0	2	2	Two brief tone pips leads to illusion of two screen flashes (annulus with checkerboard) when only one flash present
	126	Watkins	2006	Table 1 (A enhances V in general)		pooled	5	3	8	Single brief tone pip leads to illusion of single screen flash (annulus with checkerboard) when two flashes present
77	127	Watkins	2007	Fig 3 2 flashes + 1 beep illusion	10		0	1	1	Two visual flashes and single audio beep leads to the illusion of a single flash
78	128	Watson	2014a	Table 1A AV-adaptation effect (multimodal localizer)	18		0	1	1	Videos of emotional faces and voice: multisensory localizer
	129	Watson	2014a	Table 1C AV-adaptation effect, cross-modal adaptation effect		pooled	0	1	1	Videos of emotional faces and voice: crossmodal adaptation effects
79	130	Watson	2014b	Table 1 AV > baseline (Living and non-Living)	40		3	5	8	Moving objects and videos of faces with corresponding sounds: AV > baseline
	131	Watson	2014b	Table 4A integrative regions (Living and non-Living)		pooled	2	2	4	Moving objects and videos of faces with corresponding sounds: Integrative regions (AV>A,V)
	132	Watson	2014b	Table 4B integrative regions (Living and non-Living)		pooled	0	1	1	Moving objects and videos of faces with corresponding sounds: People-selective integrative region
80	133	Werner	2010	Table 1 superadditive (AV-salience effect)	21		0	3	3	Categorize movies of actions with tools or musical instruments (degraded stimuli): AV interactions both tasks
	134	Werner	2010	Table 2 AV interactions predict behavior		pooled	1	2	3	Categorize movies of actions with tools or musical instruments: AV interactions predicted by behavior
	135	Werner	2010	Table 3C superadditive AV due to task		pooled	3	0	3	Categorize movies of actions with tools or musical instruments: Subadditive AV to task
82	137	Wolf	2014	Table 1 face cartoons + phonemes	16		1	1	2	Drawing of faces with emotional expressions: Supramodal effects with emotional valence

Downloaded from https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/igab002/6103812 by West Virginia University user on 05 February 2021

Table 3. Studies included in the Incongruent category for audio-visual interaction

site meta-analyses. Results shown in Fig. 3B black hues. Refer to Tables 1 and 2 for other details.

Study #	Experiment #	First author	Year	Experimental code and abbreviated task	Subjects	Multiple experiments	Left hem foci	Right hem foci	Number of foci	Brief description of experimental paradigm
18	19	Belardinelli	2004	Table 2 AV semantic incongruent	307		55	50	105	Colored images of tools, animals, humans and semantically incongruent vs congruent sounds
8	10	Bonath	2013	bg 116 incongruent	18		1	1	2	Small checkerboards and tones: spatially incongruent vs congruent (thalamus)
12	15	Calvert	2000	Table 2 incongruent subadditive AVspeech	10		3	3	6	Speech & lower face: sub-additive AV response to incongruent AV inputs
21	33	DeHaas	2013	Table 1B V- AV incongruent	15		2	0	2	Video clips of natural scenes (animals, humans): Visual vs AV incongruent
24	38	Gonzalo	2000	Table 2 inconsistent AV	14		4	4	8	Learn novel Kanji characters & musical chords, activity increases over time to inconsistent pairings
25	40	Green	2009	Table 1 incongruent > congruent gesture-speech	16		4	5	9	Incongruent vs congruent gesture-speech
26	43	Hagan	2013	Table 2 AV emotion incongruent	18		1	5	6	Affective audio-visual speech: incongruent AV emotion vs A, V, unique ROIs over time (MEG)
31	48	Hein	2007	Fig 2A AV incongruent	18		0	2	2	Familiar animal images & incorrect (incongruent) vocalizations (dog: meow) vs correct pairs
	51	Hein	2007	Fig 3A incongruent			4	0	4	AV familiar incongruent vs unfamiliar artificial (red foci 1,5,6,9)
32	55	Hocking	2008	Table 3 incongruent simultaneous matching	18		8	10	18	Incongruent sequential AV pairs (e.g. see drum, hear bagpipes) vs congruent pairs
42	69	Matchin	2014	Table 3 MM > AV McGurk	20		7	4	11	McGurk Mismatch > AV speech integration
46	74	Murase	2008	Fig 4 discordant > concordant AVinteraction	28		1	0	1	Audiovisual speech (syllables) showing activity to discordant versus concordant stimuli: left mid-STG
48	78	Naghavi	2011	Fig 2C incongruent > congruent	30		1	2	3	B/W drawings of objects (living and non) and natural sounds (barking, piano): incongruent > congruent encoding
50	82	Naumer	2008	Fig 4 Table 2	18		1	1	2	Learn of "Freebles" and distorted sounds as incongruent > congruent pairs
52	84	Noppeny	2008	Table 2 AV incongruent > congruent	17		5	2	7	Speech sound recognition through AV priming, environmental sounds and spoken words: incongruent > congruent
68	112	Szyck	2009	Table 1 AV incongruent > AV congruent face+speech	11		7	2	9	Incongruent AV face-speech vs congruent AV face-speech
70	116	Taylor	2006	pg 8240 AV incongruent	15		1	0	1	Color photos (V), environmental sounds (A), spoken words (W): Incongruent (living objects)
72	121	Van Atteveldt	2007	Table 4 (Fig 6) active condition, incongruent	12		1	6	7	Single letters & their speech sounds (phonemes): incongruent > congruent
81	136	Willems	2007	Table 3C-D mismatch hand gestures and speech	16		2	1	3	Mismatch of hand gesture (no face) and speech versus correct

UNCORRECTED MANUSCRIPT

Table 4. Locations of significant clusters from the meta-analyses involving Congruent and Incongruent audio-visual paradigms (from Tables 2 and 3). Also indicated are major contributing studies to the ALE meta-analysis clusters, weighted centers of mass (x, y, and z) in Talairach coordinates, brain volumes (mm³), and ALE values. **(A)** single study Congruent clusters, and **(B)** single study Incongruent clusters (both corrected FWE $p < .05$); plus contrast meta-analyses maps revealing **(C)** Congruent > Incongruent and **(D)** Incongruent > Congruent audio-visual interaction sites (both uncorrected $p < .05$). The coordinates correspond to foci illustrated in Fig. 3B (black and white hues). Refer to text for other details.

Condition	Region	Major Contributing studies	x	y	z	Volume	ALE value
A. Congruent audio-visual stimuli > baseline							
1	L Posterior Superior Temporal Sulcus/Gyrus	32 (3-4,6-7,10,13,18,19,22,27,31,32,40,42,45,49,50,54,57,60,63,65-67,69,71-74,76,79,82)	-51	-36	9	4824	0.055
2	R Posterior Superior Temporal Sulcus (pSTS)	27(4,7,13,16,20,23,31,34,40,42,45,50,54,55,57,60,60,65,66,71,72,73,74,78-80)	51	-29	10	4064	0.045
3	L Inferior Frontal Cortex (posterior IFC)	4 (50,52,78,80)	-42	7	25	312	0.035
4	R Inferior Frontal Cortex (posterior IFC)	3 (27,44,50)	46	6	31	216	0.034
B. Incongruent audio-visual stimuli > baseline							
1	R Middle Frontal Gyrus / anterior IFC	3 (31,52,68)	45	14	25	320	0.024
2	L Middle Frontal Gyrus / anterior IFC	3 (8,52,72)	-40	11	29	216	0.022
C. Contrast Congruent > Incongruent audio-visual stimuli							
1	R Posterior Superior Temporal Gyrus	8 (20,40,55,60,65,66)	52	-33	13	1112	2.820
2	L Posterior Superior Temporal Gyrus		-51	-42	8	168	2.149
3	L Posterior Superior Temporal Gyrus		-49	-25	5	72	1.862
D. Contrast Incongruent > Congruent audio-visual stimuli							
1	R Inferior Frontal Cortex (Middle Frontal Gyrus)	4 (31,52,68,72)	45	13	25	416	3.540
2	L Inferior Frontal Cortex (area 9)	2 (8,72)	-41	12	28	392	2.911
3	L Inferior Frontal Cortex (area 13)	2 (42,52)	-32	20	4	56	1.932

Table 5. Studies included in the Living category for audio-visual interaction site

meta-analyses. Results shown in Fig. 3C orange hues. Refer to Tables 1-2 and text for other details.

Study #	Experiment #	First author	Year	Experimental code and abbreviated task	Subjects	Multiple experiments	Left hem foci	Right hem foci	Number of foci	Brief description of experimental paradigm
43	58				699		145	126	271	
5	6	Baumgaertner	2007	Table 3 Action > non-act sentence+video	19		3	0	3	Conjunction spoken sentences (actions>non-actions) AND videos (actions > non-actions)
6	7	Beauchamp	2004a	Fig 3J-K, Table 1 first 2 foci only	26		2	0	2	See photographs of tools, animals and hear corresponding sounds vs scrambled images and synthesized rippled sounds
7	8	Beauchamp	2004b	Expt 1 coordinates	8		1	1	2	High resolution version of 2004a study: AV tool videos vs unimodal (AV > A,V)
8	9	Belardinelli	2004	Table 1 AV semantic congruence	13		6	6	12	Colored images of tools, animals, humans and semantically congruent vs incongruent sounds
9	11	Biau	2016	Table 1A Interaction; speech synchronous	17		8	0	8	Hand gesture beats versus cartoon disc and speech interaction: synchronous vs asynchronous
11	13	Blank	2013	Fig 2	19		1	0	1	Visual-speech recognition correlated with recognition performance
16	22	Callan	2014	Table 5 AV-Audio (AV10-A10)-(AV6-A6)	16		4	4	8	Multisensory enhancement to visual speech in noise correlated with behavioral results
	23	Callan	2014	Table 6 AV - Visual only		pooled	1	1	2	Multisensory enhancement to visual speech audio-visual versus visual only
17	24	Calvert	1999	Table 1 (Fig 1)	5		3	4	7	View image of lower face and hear numbers 1 through 10 vs unimodal conditions (AV > Photos, Auditory)
18	25	Calvert	2000	Fig 2 superadditive+subadditive Avspeech	10		1	0	1	Speech & lower face: supra-additive plus sub-additive effects (AV-congruent > A,V > AV-incongruent)
	26	Calvert	2000	Table 1 supraditive AVspeech		pooled	4	5	9	Speech & lower face: supra-additive AV enhancement
20	31	Calvert	2003	Table 2A (Fig 3 blue)	8		13	8	21	Speech and lower face: Moving dynamic speech (phonemes) versus stilled speech frames
21	32	DeHaas	2013	Table 1A AVcong - Visual	15		3	3	6	Video clips of natural scenes (animals, humans): AV congruent vs Visual
22	34	Erickson	2014	Table 1A Congruent AV speech	10		2	2	4	McGurk effect (phonemes): congruent AV speech: AV > A and AV > V
	35	Erickson	2014	Table 1B McGurk speech		pooled	2	0	2	McGurk speech effect (phonemes)
23	36	Ethofer	2013	Table 1C emotion	23		1	2	3	Audiovisual emotional face-voice integration
25	41	Green	2009	Table 4A Congruent gesture-speech > gesture or speech	16		1	0	1	Congruent gesture-speech vs gesture with unfamiliar speech and with familiar speech
27	44	Hasegawa	2004	Table 1A (well trained piano) AV induced by V-only	26		12	6	18	Piano playing: well trained pianists, mapping hand movements to sequences of sound
29	46	He	2015	Table 3C AV speech foreign (left MTG focus)	20		1	0	1	Intrinsically meaningful gestures with German speech: Gesture-German > Gesture-Russian, German speech only
30	47	He	2018	Table 2 gestures & speech integration (left MTG)	20		1	0	1	Gesture-speech integration: Bimodal speech-gesture vs unimodal gesture with foreign speech and vs unimodal
31	50	Hein	2007	Fig 2C pSTS, pSTG, mSTG AV-cong	18		0	3	3	Familiar animal images & correct vocalizations (dog: woof-woof)
32	54	Hocking	2008	pg 2444 verbal	18		2	0	2	(pSTS mask) Color photos, written names, auditory names, environmental sounds conceptually matched "amodal"
35	58	James	2011	Table 1A bi-modal (vs scrambled)	12		4	2	6	Video of human manual actions (e.g. sawing): Auditory and Visual intact versus scrambled, AV event selectivity
36	59	Jessen	2015	Table 1A emotion > neutral AV enhanced	17		1	1	2	Emotional multisensory whole body and voice expressions: AV emotion (anger and fear) > neutral expressions
	60	Jessen	2015	Table 1D fear > neutral AV enhanced		pooled	2	1	3	Emotional multisensory whole body and voice expressions: AV fear > neutral expressions
37	61	Jola	2013	Table 1C AVcondition dance	12		3	3	6	Viewing unfamiliar dance performance (tells a story by gesture) with vs without music: using intersubject correlation
38	62	Kim	2015	Table 2A AV>C speech semantic match	15		2	0	2	Moving audio-visual speech perception vs white noise and unopened mouth movements
39	63	Kircher	2009	Figure 3B gesture related activation increase	14		3	1	4	Bimodal gesture-speech vs gesture and vs speech
40	64	Kreifelts	2007	Table 1 voice-face emotion	24		1	2	3	Facial expression & intonated spoken words, judge emotion expressed (AV > A,V; p<0.05 only)
	65	Kreifelts	2007	Table 5 AV increase effective connectivity		pooled	2	4	6	Increased effectiveness connectivity with pSTS and thalamus during AV integration of non-verbal emotional information
42	67	Matchin	2014	Table 1 AV > Aud only (McGurk)	20		2	7	9	McGurk audio-visual speech: AV > A only
	68	Matchin	2014	Table 2 AV > Video only		pooled	9	6	15	McGurk audio-visual speech: AV > V only
45	73	Muller	2012	Table S1 effective connectivity changes	27		4	3	7	Emotional facial expression (groaning, laughing) AV integration and gating of information
49	79	Nath	2012	pg 784	14		1	0	1	McGurk effect (phonemes): congruent AV speech correlated with behavioral percept

54	87	Ogawa	2013b	Table 1 3D > 2D and surround > monaural effects	16		3	4	7	Cinematic 3D > 2D video and surround sound > monaural while watching a movie ("The Three Musketeers")
55	88	Okada	2013	Table 1 AV > A	20		5	4	9	Video of AV > A speech only
56	89	Olson	2002	Table 1A synchronized AV > static Vis-only	10		7	4	11	Whole face video & heard words: Synchronized AV vs static V
	90	Olson	2002	Table 1C synchronized AV > desynchronized AV speech		pooled	2	0	2	Whole face video & heard words: Synchronized vs desynchronized
60	96	Robins	2008	Table 4A (Fig 5) AV integration & emotion	5		1	4	5	AV faces and spoken sentences expressing fear or neutral valence: AV integration (AV > A, V conditions)
	97	Robins	2008	Table 4B emotion effects		pooled	2	0	2	AV faces and spoken sentences expressing fear or neutral valence: Emotional AV-fear > AV-neutral
	98	Robins	2008	Table 4C (Fig 5) fearful AV integration		pooled	1	5	6	AV faces and spoken sentences expressing fear or neutral valence: Fearful-only AV integration
	99	Robins	2008	Table 4D AV-only emotion		pooled	1	3	4	AV faces and spoken sentences expressing fear or neutral valence: AV-only emotion
61	100	Scheef	2009	Table 1 cartoon jump + boing	16		1	2	3	Video of cartoon person jumping & "sonification" of a tone, learn correlated pairings: AV-V and AV-A conjunction
63	104	Sekiyama	2003	Table 3 (fMRI nAV-AV)	8		1	0	1	AV speech, McGurk effect with phonemes (ba, da, ga) and noise modulation: noise-AV > AV (fMRI)
	105	Sekiyama	2003	Table 4 (PET nAV-AV)		pooled	1	3	4	AV speech, McGurk effect with phonemes (ba, da, ga) and noise modulation: noise-AV > AV (PET)
64	106	Sestieri	2006	Table 1 (Fig 3), AV location match vs semantic	10		2	5	7	B/W images (animal, weapons) & environmental sounds: Match location > recognition
65	108	Stevenson	2009	Table 1B AVtools > AVspeech	11		1	1	2	Hand tools in use video: inverse effectiveness (degraded AV tool > AV speech)
	109	Stevenson	2009	Table 1C (Fig 8) AVspeech > AVtools		pooled	1	1	2	Face & speech video: inverse effectiveness (degraded AV speech > AV tool use)
66	110	Straube	2011	Table 3A-B iconic/metaphoric speech-gestures vs speech, gestu	16		2	2	4	Integration of iconic and Metaphoric speech-gestures vs speech and gesture
67	111	Straube	2014	p939 Integration foci	16		3	0	3	Integration of iconic hand gesture-speech > unimodal speech and unimodal gesture (healthy control group)
75	124	Von Kriegstein	2006	Fig 4B after > before voice-face	14		0	4	4	Face and object photos with voice and other sounds: Voice-Face association learning
78	128	Watson	2014a	Table 1A AV-adaptation effect (multimodal localizer)	18		0	1	1	Videos of emotional faces and voice: multisensory localizer
	129	Watson	2014a	Table 1C AV-adaptation effect, cross-modal adaptation effect		pooled	0	1	1	Videos of emotional faces and voice: crossmodal adaptation effects
79	132	Watson	2014b	Table 4B integrative regions (Living and non-Living)	40		0	1	1	Moving objects and videos of faces with corresponding sounds: People-selective integrative region
80	133	Werner	2010	Table 1 superadditive (AV-salience effect)	21		0	3	3	Categorize movies of actions with tools or musical instruments (degraded stimuli); AV interactions both tasks
	134	Werner	2010	Table 2 AV inteactions predict behavior		pooled	1	2	3	Categorize movies of actions with tools or musical instruments; AV interactions predicted by behavior
	135	Werner	2010	Table 3C superadditive AV due to task		pooled	3	0	3	Categorize movies of actions with tools or musical instruments; Subadditive AV to task
82	137	Wolf	2014	Table 1 face cartoons + phonemes	16		1	1	2	Drawing of faces with emotional expressions: Supramodal effects with emotional valence

UNCORRECTED MANUSCRIPT

Downloaded from https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/igab002/6103812 by West Virginia University user on 05 February 2021

Table 6. Studies included in the Non-living category for audio-visual interaction site meta-analyses. Results shown in Fig. 3C cyan. Refer to Table 1 and text for other details.

Study #	Experiment #	First author	Year	Experimental code and abbreviated task	# Subjects	Multiple experiments	Left hem foci	Right hem foci	Number of foci	Brief description of experimental paradigm
15	25				187		93	93	186	
1	1	Adams	2002	Expt 1 Table 3 A+V (aud coords only)	12		5	1	6	A and V commonly showing subordinate > basic object name verification (words with pictures or environmental sounds)
2	2	Alink	2008	Table 1c spheres move to drum sounds	10		4	6	10	Visual spheres and drum sounds moving; crossmodal dynamic capture vs conflicting motion
4	4	Baumann	2007	Table 1B coherent V+A vs A	12		2	1	3	Visual dots 16% coherent motion & in-phase acoustic noise > stationary acoustic sound
	5	Baumann	2007	Table 2B		pooled	15	12	27	Moving acoustic noise & visual dots 16% in-phase coherent > random dot motion
12	14	Bonath	2013	pg 116 congruent thalamus	18		1	0	1	Small checkerboards and tones; spatially congruent vs incongruent (thalamus)
13	16	Bonath	2014	Table 1A illusory vs not	20		1	5	6	Small checkerboards and tones; temporal > spatial congruence
	17	Bonath	2014	Table 1B synchronous > no illusion		pooled	3	0	3	Small checkerboards and tones; spatial > temporal congruence
14	18	Bushara	2001	Table 1A (Fig 2) AV-Control	12		1	3	4	Tones (100 ms) & colored circles synchrony: detect Auditory then Visual presentation vs Control
	19	Bushara	2001	Table 1B (VA-C) five coords		pooled	2	3	5	Tones (100 ms) & colored circles synchrony: detect Visual then Auditory presentation vs Control
	20	Bushara	2001	Table 2A interact w/ Rt Insula		pooled	2	4	6	Tones & colored circles: correlated functional connections with (and including) the right insula
15	21	Bushara	2003	Table 2A collide > pass, strong A-V interact	7		5	3	8	Tone and two visual bars moving: Tone synchrony induce perception they collide (AV interaction) vs pass by
19	28	Calvert	2001	Table 2 superadditive & response depression	10		4	11	15	B/W visual checkerboard reversing & white noise bursts: Synchronous vs not; supraditive and response depression
	29	Calvert	2001	Table 3A superadditive only		pooled	6	4	10	B/W visual checkerboard reversing & white noise bursts: Synchronous vs not; supraditive only
	30	Calvert	2001	Table 3B response depression only		pooled	3	4	7	B/W visual checkerboard reversing & white noise bursts: Synchronous vs not; response depression only
33	56	Hove	2013	pg 316 AV interaction putamen	14		0	1	1	Interaction between (beep > flash) vs (siren > moving bar); left putamen focus
41	66	Lewis	2000	Table 1	7		2	3	5	Compare speed of tone sweeps to visual dot coherent motion: Bimodal vs unimodal
44	71	Meyer	2007	Table 3 paired A+V vs null	16		3	3	6	Paired screen red flashes with phone ring: paired V (conditioned stimulus) and A (unconditioned) vs null events
	72	Meyer	2007	Table 4 CS+ learned AV association with V-only		pooled	4	6	10	Paired screen flashes with phone ring: View flashes after post-conditioned vs null events
53	86	Ogawa	2013a	Table 1 (pg 162 data)	13		1	0	1	AV congruency of pure tone and white dots moving on screen (area left V3A)
69	113	Tanabe	2005	Table 1A AV; A then V; not VA	15		10	10	20	Amorphous texture patterns & modulated white noises: Activation during learning delay period (AV)
	114	Tanabe	2005	Table 2A+2B (Fig 5a) AV and VA		pooled	5	6	11	Amorphous texture patterns & modulated white noises: changes after feedback learning (AV and VA)
	115	Tanabe	2005	Table 3A+3B (Fig 6) AV and VA; delay period		pooled	9	1	10	Amorphous texture patterns & modulated white noises: sustained activity throughout learning (AV and VA)
76	125	Watkins	2006	Fig 4 illusory multisensory interaction	11		0	2	2	Two brief tone pips leads to illusion of two screen flashes (annulus with checkerboard) when only one flash present
	126	Watkins	2006	Table 1 (A enhances V in general)		pooled	5	3	8	Single brief tone pip leads to illusion of single screen flash (annulus with checkerboard) when two flashes present
77	127	Watkins	2007	Fig 3 2 flashes + 1 beep illusion	10		0	1	1	Two visual flashes and single audio beep leads to the illusion of a single flash

UNCORRECTED MANUSCRIPT

Downloaded from https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/igab002/6103812 by West Virginia University user on 05 February 2021

Table 7. Locations of significant clusters from the meta-analyses involving Congruent and Incongruent audio-visual experimental paradigms (from Tables 5 and 6), indicating major contributing studies to the ALE meta-analysis clusters, weighted centers of mass (x, y, and z) in Talairach coordinates, brain volumes (mm³), and ALE values. **(A)** Single study ALE maps for Living (corrected FWE $p < .05$) and **(B)** Non-living audio-visual interaction sites (corrected FWE $p < .05$); plus contrast meta-analyses maps revealing **(C)** Living > Non-living, and **(D)** Non-living > Living audio-visual interaction-sites (uncorrected $p < .05$). The coordinates correspond to foci illustrated in Fig. 3C (orange and cyan).

Condition	Region	Major Contributing studies	x	y	z	Volume	ALE value
A. Living audio-visual stimuli > baseline							
1	L Superior Temporal Sulcus, posterior (pSTS)	8 (6,18,32,40,63,65-67)	-50	-51	10	1448	0.042
2	R Superior Temporal Sulcus	8 (20,31,40,60,65,66,78,79)	48	-37	12	1280	0.035
3	R superior Temporal Gyrus (pSTG)	3 (40,42,45)	55	-19	7	256	0.025
4	L Superior Temporal Gyrus	2 (45,49)	-53	-23	7	144	0.024
B. Non-Living audio-visual stimuli > baseline							
1	R Anterior Insula	1 (44)	31	19	6	32	0.019
C. Living > Non-living audio-visual stimuli							
1	L Posterior Superior Temporal Sulcus	2 (40,66)	-50	-52	12	408	2.054
2	R Posterior Superior Temporal Sulcus	1 (66)	51	-35	12	48	1.779
D. Non-living > Living audio-visual stimuli							
1	R Anterior Insula	1 (44)	31	19	6	32	3.195

Table 8. Studies included in the Vocalizations category for audio-visual interaction site meta-analyses. Results shown in Fig. 3D red hues. Refer to Tables 1-2 and text for other details.

Study #	Experiment #	First author	Year	Experimental code and abbreviated task	# Subjects	Multiple experiments	Left hem foci	Right hem foci	Number of foci	Brief description of experimental paradigm
40	57				647		146	117	263	
3	3	Balk	2010	Fig 2 asynchronous vs simultaneous	14		2	1	3	Natural asynchronous vs simultaneous AV speech synchrony (included both contrasts as interaction effects)
5	6	Baumgaertner	2007	Table 3 Action > non-act sentence+video	19		3	0	3	Conjunction spoken sentences (actions-non-actions) AND videos (actions > non-actions)
9	11	Biau	2016	Table 1A Interaction; speech synchronous	17		8	0	8	Hand gesture beats versus cartoon disc and speech interaction: synchronous vs asynchronous
11	13	Blank	2013	Fig 2	19		1	0	1	Visual-speech recognition correlated with recognition performance
16	22	Callan	2014	Table 5 AV-Audio (AV10-A10)-(AV6-A6)	16		4	4	8	Multisensory enhancement to visual speech in noise correlated with behavioral results
	23	Callan	2014	Table 6 AV - Visual only		pooled	1	1	2	Multisensory enhancement to visual speech audio-visual versus visual only
17	24	Calvert	1999	Table 1 (Fig 1)	5		3	4	7	View image of lower face and hear numbers 1 through 10 vs unimodal conditions (AV > Photos, Auditory)
18	25	Calvert	2000	Fig 2 superadd+subadd AVspeech	10		1	0	1	Speech & lower face: supra-additive plus sub-additive effects (AV > A and AV > V)
	26	Calvert	2000	Table 1. supradd AVspeech		pooled	4	5	9	Speech & lower face: supra-additive AV enhancement
20	31	Calvert	2003	Table 2A (Fig 3 blue)	8		13	8	21	Speech and lower face: Moving dynamic speech (phonemes) versus stilled speech frames
22	34	Erickson	2014	Table 1A Congruent AV speech	10		2	2	4	McGurk effect (phonemes): congruent AV speech: AV > A and AV > V
	35	Erickson	2014	Table 1B McGurk speech		pooled	2	0	2	McGurk speech effect (phonemes)
23	36	Ethofer	2013	Table 1C emotion	23		1	2	3	Audiovisual emotional face-voice integration
25	41	Green	2009	Table 4A Congruent gesture-speech > gesture or speech	16		1	0	1	Congruent gesture-speech vs gesture with unfamiliar speech and with familiar speech
26	42	Hagan	2013	Table 1 AV emotion, novel over time	18		5	3	8	Affective audio-visual speech: congruent AV emotion vs A, V; unique ROIs over time (MEG)
28	45	Hashimoto	2004	Table 1G (Fig 4B, red) Learning Hangul letters to sounds	12		2	1	3	Unfamiliar Hangul letters & nonsense words, learn speech vs tone/noise pairings
29	46	He	2015	Table 3C AV speech foreign (left MTG focus)	20		1	0	1	Intrinsically meaningful gestures with German speech: Gesture-German > Gesture-Russian, German speech only
30	47	He	2018	Table 2 gestures & speech integration	20		1	0	1	Gesture-speech integration: Bimodal speech-gesture vs unimodal gesture with foreign speech and vs unimodal
31	50	Hein	2007	Fig 2C pSTS, pSTG, mSTG AV-cong	18		0	3	3	Familiar animal images & correct vocalizations (dog: woof-woof)
	52	Hein	2007	Fig 3B Foci 2, 3, 4 (blue) artificial/non-living		pooled	3	0	3	Visual "Fribbles" & backward/underwater distorted animal sounds, learn pairings (blue foci 2,3,4)
	53	Hein	2007	Fig 3C congruent living (green)		pooled	3	0	3	Familiar congruent living vs artificial AV object features and animal sounds (green foci 7, 8, 10)
36	59	Jessen	2015	Table 1A emotion > neutral AV enhanced	17		1	1	2	Emotional multisensory whole body and voice expressions: AV emotion (anger and fear) > neutral expressions
	60	Jessen	2015	Table 1D fear > neutral AV enhanced		pooled	2	1	3	Emotional multisensory whole body and voice expressions: AV fear > neutral expressions
38	62	Kim	2015	Table 2A AV>C speech semantic match	15		2	0	2	Moving audio-visual speech perception vs white noise and unopened mouth movements (AV > C)
39	63	Kircher	2009	Figure 3B: gesture related activation increase	14		3	1	4	Bimodal gesture-speech vs gesture and vs speech
40	64	Kreifelts	2007	Table 1 voice-face emotion	24		1	2	3	Facial expression & intonated spoken words, judge emotion expressed (AV > A,V; p<0.05 only)
	65	Kreifelts	2007	Table 5 AV increase effective connectivity		pooled	2	4	6	Increased effectiveness connectivity with pSTS and thalamus during AV integration of non-verbal emotional information
42	67	Matchin	2014	Table 1 AV > Aud only (McGurk)	20		2	7	9	McGurk audio-visual speech: AV > A only
	68	Matchin	2014	Table 2 AV > Video only		pooled	9	6	15	McGurk audio-visual speech: AV > V only
45	73	Muller	2012	Table S1 effective connectivity changes	27		4	3	7	Emotional facial expression (groaning, laughing) AV integration and gating of information
49	79	Nath	2012	pg 784	14		1	0	1	McGurk effect (phonemes): congruent AV speech correlated with behavioral percept
50	80	Naumer	2008	Fig 2 Table 1A max contrast	18		8	6	14	Images of "Fribbles" & learned artificial sounds (underwater animal vocals): post training vs max contrast
	81	Naumer	2008	Fig 3 Table 1B pre-post		pooled	5	6	11	Images of "Fribbles" & learned corresponding artificial sounds: Post- vs Pre-training session

55	88	Okada	2013	Table 1 AV > A	20		5	4	9	Video of AV > A speech only
56	89	Olson	2002	Table 1A synchronized AV > static Vis-only	10		7	4	11	Whole face video & heard words: Synchronized AV vs static V
	90	Olson	2002	Table 1C synchronized AV > desynchronized AV speech		pooled	2	0	2	Whole face video & heard words: Synchronized vs desynchronized
58	93	Raij	2000	Table 1B letters & speech sounds	9		2	3	5	Integration of visual letters and corresponding auditory phonetic expressions (MEG study) AV vs (A + V)
60	95	Robins	2008	Table 2 (Fig 2) AV integration (AV > A and AV > V)	10		2	1	3	Face speaking sentences: angry, fearful, happy, neutral (AV > A, V)
	96	Robins	2008	Table 4A (Fig 5) AV integration & emotion	5		1	4	5	AV faces and spoken sentences expressing fear or neutral valence: AV integration (AV > A, V conditions)
	97	Robins	2008	Table 4B emotion effects		pooled	2	0	2	AV faces and spoken sentences expressing fear or neutral valence: Emotional AV-fear > AV-neutral
	98	Robins	2008	Table 4C (Fig 5) fearful AV integration		pooled	1	5	6	AV faces and spoken sentences expressing fear or neutral valence: Fearful-only AV integration
	99	Robins	2008	Table 4D AV-only emotion		pooled	1	3	4	AV faces and spoken sentences expressing fear or neutral valence: AV-only emotion
63	104	Sekiyama	2003	Table 3 (fMRI nAV-AV)	8		1	0	1	AV speech, McGurk effect with phonemes (ba, da, ga) and noise modulation: noise-AV > AV (fMRI)
	105	Sekiyama	2003	Table 4 (PET nAV-AV)		pooled	1	3	4	AV speech, McGurk effect with phonemes (ba, da, ga) and noise modulation: noise-AV > AV (PET)
65	109	Stevenson	2009	Table 1C (Fig 8) AVspeech > AVtools	11		1	1	2	Face & speech video: inverse effectiveness (degraded AV speech > AV tool use)
66	110	Straube	2011	Table 3A-B iconic/metaphoric speech-gestures vs speech, gestu	16		2	2	4	Integration of iconic hand gesture-speech vs speech and gesture
67	111	Straube	2014	p939 Integration foci	16		3	0	3	Integration of iconic hand gesture-speech > unimodal speech and unimodal gesture (healthy control group)
71	118	Van Atteveldt	2004	Table 1a letters & speech sounds	16		3	1	4	Familiar letters & their speech sounds: Congruent vs not and Bimodal vs Unimodal
72	119	Van Atteveldt	2007	Table 2A+B (Fig 2)	12		3	2	5	Single letters & their speech sounds (phonemes): Congruent > Incong; Passive perception, blocked and event-related
	120	Van Atteveldt	2007	Table 3 (Fig 2) passive		pooled	1	1	2	Single letters & their speech sounds (phonemes): Congruent > Incong, active perception task
73	122	Van Atteveldt	2010	Table 1B STS: specific adaptation congruent > incong	16		3	1	4	Letter and speech sound pairs (vowels, consonants): Specific adaptation effects
74	123	Van der Wyk	2010	Table 2 AV interaction effects oval/circles+speech/nonspeech	16		3	3	6	Geometric shape modulate with speech (sentences)
75	124	Von Kriegstein	2006	Fig 4B after > before voice-face	14		0	4	4	Face and object photos with voice and other sounds: Voice-Face association learning
78	128	Watson	2014a	Table 1A AV-adaptation effect (multimodal localizer)	18		0	1	1	Videos of emotional faces and voice: multisensory localizer
	129	Watson	2014a	Table 1C AV-adaptation effect, cross-modal adaptation effect		pooled	0	1	1	Videos of emotional faces and voice: crossmodal adaptation effects
79	132	Watson	2014b	Table 4B integrative regions (Living and non-Living)	40		0	1	1	Moving objects and videos of faces with corresponding sounds: People-selective integrative region
82	137	Wolf	2014	Table 1 face cartoons + phonemes	16		1	1	2	Drawing of faces with emotional expressions: Supramodal effects with emotional valence

UNCORRECTED MANUSCRIPT

Downloaded from https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/igab002/6103812 by West Virginia University user on 05 February 2021

Table 9. Studies included in the Actions category for audio-visual interaction site meta-analyses. Results shown in Fig. 3D yellow. Refer to Tables 1-2 and text for other details.

Study #	Experiment #	First author	Year	Experimental code and abbreviated task	# Subjects	Multiple experiments	Left hem foci	Right hem foci	Number of foci	Brief description of experimental paradigm
13	19				205		50	50	100	
6	7	Beauchamp	2004a	Fig 3J-K, Table 1 first 2 foci only	26		2	0	2	See photographs of tools, animals and hear corresponding sounds vs scrambled images and synthesized rippled sounds
7	8	Beauchamp	2004b	Expt 1 coordinates	8		1	1	2	High resolution version of 2004a study: AV tool Videos vs unimodal (AV > A,V)
8	9	Belardinelli	2004	Table 1 AV semantic congruence	13		6	6	12	Colored images of tools, animals, humans and semantically congruent vs incongruent sounds
27	44	Hasegawa	2004	Table 1A (well trained piano) AV induced by V-only	26		12	6	18	Piano playing: well trained pianists, mapping hand movements to sequences of sound
35	58	James	2011	Table 1A bi-modal (vs scrambled)	12		4	2	6	Video of human manual actions (e.g. sawing): Auditory and Visual intact versus scrambled. AV event selectivity
37	61	Jola	2013	Table 1C AV condition dance	12		3	3	6	Viewing unfamiliar dance performance (tells a story by gesture) with vs without music: using intersubject correlation
47	75	Naghavi	2007	Fig 1C	23		0	3	3	B/W pictures (animals, tools, instruments, vehicles) & their sounds: Cong vs Incong
57	91	Plank	2012	pg 803 AV congruent effect	15		0	1	1	AV spatially congruent > semantically matching images of natural objects and associated sounds (right STG)
	92	Plank	2012	Table 2A spatially congruent-baseline		pooled	5	5	10	Images of natural objects and associated sounds, spatially congruent vs baseline
61	100	Scheef	2009	Table 1 cartoon jump + boing	16		1	2	3	Video of cartoon person jumping & "sonification" of a tone, learn correlated pairings: AV-V and AV-A conjunction
62	101	Schmid	2011	Table 2E A effect V (Living & non-living, pictures)	12		3	4	7	Environmental sounds & matching pictures: reduced activity by A
	102	Schmid	2011	Table 2F V competition effect A (reduced activity by a visual object)		pooled	2	2	4	Environmental sounds & matching pictures: reduced activity by V
	103	Schmid	2011	Table 2G AV crossmodal interaction x auditory attention		pooled	2	3	5	Environmental sounds and matching pictures: cross-modal interaction and auditory attention
64	106	Sestieri	2006	Table 1 (Fig 3), AV location match vs semantic	10		2	5	7	B/W images (animal, weapons) & environmental sounds: Match location > recognition
	107	Sestieri	2006	Table 2 AV semantic recognition vs localization		pooled	2	1	3	B/W pictures and environmental sounds: congruent semantic recognition > localization task
65	108	Stevenson	2009	Table 1B AVtools > AVspeech	11		1	1	2	Hand tools in use video: inverse effectiveness (degraded AV tool > AV speech)
80	133	Werner	2010	Table 1 superadditive (AV-salience effect)	21		0	3	3	Categorize movies of actions with tools or musical instruments (degraded stimuli); AV interactions both tasks
	134	Werner	2010	Table 2 AV interactions predict behavior		pooled	1	2	3	Categorize movies of actions with tools or musical instruments; AV interactions predicted by behavior
	135	Werner	2010	Table 3C superadditive AV due to task		pooled	3	0	3	Categorize movies of actions with tools or musical instruments; Subadditive AV to task

UNCORRECTED MANUSCRIPT

Downloaded from https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/igab002/6103812 by West Virginia University user on 05 February 2021

Table 10. Studies included in the Emotional audio-visual interaction site meta-

analyses. Most of these studies used vocalizations as auditory stimuli, and thus was included as a subset of the congruent vocalization category with results shown in Fig. 3D violet hues. Refer to Tables 1-2 and text for other details.

Study #	Experiment #	First author	Year	Experimental code and abbreviated task	# Subjects	Multiple experiments	Left hem foci	Right hem foci	Number of foci	Brief description of experimental paradigm
9	13	First author			160		24	29	53	Audiovisual emotional face-voice integration
23	36	Ethofer	2013	Table 1C emotion	23		1	2	3	Affective audio-visual speech: congruent AV emotion vs A, V; unique ROIs over time (MEG)
26	42	Hagan	2013	Table 1 AV emotion, novel over time	18		5	3	8	Emotional multisensory whole body and voice expressions: AV emotion (anger and fear) > neutral expressions
36	59	Jessen	2015	Table 1A emotion > neutral AV enhanced	17		1	1	2	Emotional multisensory whole body and voice expressions: AV fear > neutral expressions
	60	Jessen	2015	Table 1D fear > neutral AV enhanced		pooled	2	1	3	Viewing unfamiliar dance performance (tells a story by gesture) with vs without music: using intersubject correlation
37	61	Jola	2013	Table 1C AVcondition dance	12		3	3	6	Facial expression & intonated spoken words, judge emotion expressed (AV > A,V: p<0.05 only)
40	64	Kreifelts	2007	Table 1 voice-face emotion	24		1	2	3	Increased effectiveness connectivity with pSTS and thalamus during AV integration of non-verbal emotional information
	65	Kreifelts	2007	Table 5 AV increase effective connectivity		pooled	2	4	6	Emotional facial expression (groaning, laughing) AV integration and gating of information
45	73	Muller	2012	Table S1 effective connectivity changes	27		4	3	7	AV faces and spoken sentences expressing fear or neutral valence: Emotional AV-fear > AV-neutral
60	97	Robins	2008	Table 4B emotion effects	5		2	0	2	AV faces and spoken sentences expressing fear or neutral valence: Fearful-only AV integration
	98	Robins	2008	Table 4C (Fig 5) fearful AV integration		pooled	1	5	6	AV faces and spoken sentences expressing fear or neutral valence: AV-only emotion
	99	Robins	2008	Table 4D AV-only emotion		pooled	1	3	4	Videos of emotional faces and voice: crossmodal adaptation effects
78	129	Watson	2014a	Table 1C AV-adaptation effect, cross-modal adaptation effect	18		0	1	1	Drawing of faces with emotional expressions: Supramodal effects with emotional valence
82	137	Wolf	2014	Table 1 face cartoons + phonemes	16		1	1	2	

UNCORRECTED MANUSCRIPT

Downloaded from https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/igab002/6103812 by West Virginia University user on 05 February 2021

Table 11. Locations of significant clusters from the meta-analyses involving Vocalizations and Non-vocal audio-visual experimental paradigms (from Tables 8, 9 and 10), indicating major contributing studies to the ALE meta-analysis clusters, weighted centers of mass (x, y, and z) in Talairach coordinates, brain volumes (mm³), and ALE values. (A) Single study ALE maps for Vocalizations (corrected FWE $p < .05$) and (B) Action stimuli (corrected FWE $p < .05$), plus (C) contrast maps revealing interaction sites involving Vocalization $>$ Action and (D) Action $>$ Vocalization auditory stimuli (both uncorrected $p < .05$). A subset of the Vocal/Living audio-visual stimuli also entailed (E) emotionally valent audio-visual stimuli, which was conducted as a single study ALE map (corrected FWE $p < .05$). TTG = transverse temporal gyrus (aka HG = Heschl's gyrus). The coordinates correspond to foci illustrated in Figure 3D (red, yellow and violet hues).

Condition	Region	Major Contributing studies	x	y	z	Volume	ALE value
A. Vocal audio-visual stimuli > baseline							
1	R Superior Temporal Sulcus	19(20,23,26,31,40,42,45,50,55,58,60,60,65,66,71,72,74,78,79)	50	-32	11	3040	0.041
2	L Superior Temporal Sulcus (posterior), BA 22	9(18,22,40,63,66,67,71,73,74)	-54	-47	11	1328	0.034
3	L Superior Temporal Sulcus, BA 41	7(3,22,31,42,45,50,72)	-49	-21	7	1200	0.035
4	L Superior Temporal Gyrus (posterior), BA 41	4(45,50,60,65)	-47	-37	11	376	0.030
B. Non-vocal (living) audio-visual stimuli > baseline							
1	L Fusiform Gyrus (inferior-medial)	1(62)	-28	-54	-14	8	0.017
C. Vocal > Non-vocal audio-visual stimuli							
	R Posterior Superior Temporal Sulcus	7(31,40,60,65,66,78,79)	46	-37	13	976	2.530
	R Posterior Superior Temporal Gyrus		54	-26	8	8	1.672
D. Non-vocal > Vocal audio-visual stimuli							
1	L Fusiform Gyrus (inferior-medial)	1(62)	-28	-54	-14	8	2.400
E. Emotionally valent (mostly vocal) > Non-emotional stimuli							
	R Posterior Superior Temporal Gyrus	3(26,37,45)	58	-21	8	152	2.391

Table 12. Studies included in the Dynamic-visual stimuli category for audio-visual interaction site meta-analyses. Results shown in Fig. 3E blue. Refer to Tables 1-2 and text for other details.

Study #	Experiment #	First author	Year	Experimental code and abbreviated task	# Subjects	Multiple experiments	Left hem foci	Right hem foci	Number of foci	Brief description of experimental paradigm
43	62				682		177	148	325	Visual spheres and drum sounds moving: crossmodal dynamic capture vs conflicting motion
2	2	Alink	2008	Table 1c spheres move to drum sounds	10		4	6	10	Natural asynchronous vs simultaneous AV speech synchrony (included both contrasts as interaction effects)
3	3	Balk	2010	Fig 2 asynchronous vs simultaneous	14		2	1	3	Visual dots 16% coherent motion & in-phase acoustic noise > stationary acoustic sound
4	4	Baumann	2007	Table 1B coherent V+A vs A	12		2	1	3	Moving acoustic noise & visual dots 16% in-phase coherent > random dot motion
	5	Baumann	2007	Table 2B		pooled	15	12	27	Conjunction spoken sentences (actions>non-actions) AND videos (actions>non-actions)
5	6	Baumgaertner	2007	Table 3 Action > non-act sentence+video	19		3	0	3	High resolution version of 2004a study: AV tool videos vs unimodal (AV > A,V)
7	8	Beauchamp	2004b	Expt 1 coordinates	8		1	1	2	Hand gesture beats versus cartoon disc and speech interaction: synchronous vs asynchronous
9	11	Biau	2016	Table 1A Interaction; speech synchronous	17		8	0	8	Visual-speech recognition correlated with recognition performance
11	13	Blank	2013	Fig 2	19		1	0	1	Tone and two visual bars moving: Tone synchrony induce perception they collide (AV interaction) vs pass by
15	21	Bushara	2003	Table 2A collide > pass, strong A-V interact	7		5	3	8	Multisensory enhancement to visual speech in noise correlated with behavioral results
16	22	Callan	2014	Table 5 AV-Audio (AV10-A10)-(AV6-A6)	16		4	4	8	Multisensory enhancement to visual speech audio-visual versus visual only
	23	Callan	2014	Table 6 AV - Visual only		pooled	1	1	2	Speech & lower face: supra-additive plus sub-additive effects (AV-congruent > A,V > AV-incongruent)
18	25	Calvert	2000	Fig 2 superadd+subadd AVspeech	10		1	0	1	Speech & lower face: supra-additive AV enhancement
	26	Calvert	2000	Table 1. supradd AVspeech		pooled	4	5	9	Speech and lower face: Moving dynamic speech (phonemes) versus stilled speech frames
20	31	Calvert	2003	Table 2A (Fig 3 blue)	8		13	8	21	Video clips of natural scenes (animals, humans): AV congruent vs Visual
21	32	DeHaas	2013	Table 1A AVcong - Visual	15		3	3	6	McGurk effect (phonemes): congruent AV speech: AV > A and AV > V
22	34	Erickson	2014	Table 1A Congruent AV speech	10		2	2	4	McGurk speech effect (phonemes)
	35	Erickson	2014	Table 1B McGurk speech		pooled	2	0	2	Audiovisual emotional face-voice integration
23	36	Ethofer	2013	Table 1C emotion	23		1	2	3	Congruent gesture-speech vs gesture with unfamiliar speech and with familiar speech
25	41	Green	2009	Table 4A Congruent gesture-speech > gesture or speech	16		1	0	1	Affective audio-visual speech: congruent AV emotion vs A, V; unique ROIs over time (MEG)
26	42	Hagan	2013	Table 1 AV emotion, novel over time	18		5	3	8	Piano playing: well trained pianists, mapping hand movements to sequences of sound
27	44	Hasegawa	2004	Table 1A (well trained piano) AV induced by V-only	26		12	6	18	Intrinsically meaningful gestures with German speech: Gesture-German > Gesture-Russian, German speech only
29	46	He	2015	Table 3C AV speech foreign (left MTG focus)	20		1	0	1	Gestures-speech integration: Bimodal speech-gesture vs unimodal gesture with foreign speech and vs unimodal
30	47	He	2018	Table 2. GSI, left MTG, gestures & speech integration	20		1	0	1	Video of human manual actions (e.g. sawing): Auditory and Visual intact versus scrambled, AV event selectivity
35	58	James	2011	Table 1A bi-modal (vs scrambled)	12		4	2	6	Emotional multisensory whole body and voice expressions: AV emotion (anger and fear) > neutral expressions
36	59	Jessen	2015	Table 1A emotion > neutral AV enhanced	17		1	1	2	Emotional multisensory whole body and voice expressions: AV fear > neutral expressions
	60	Jessen	2015	Table 1D fear > neutral AV enhanced		pooled	2	1	3	Viewing unfamiliar dance performance (tells a story by gesture) with vs without music: using intersubject correlation
37	61	Jola	2013	Table 1C AVcondition dance	12		3	3	6	Moving audio-visual speech perception vs white noise and unopened mouth movements
38	62	Kim	2015	Table 2A AV>C speech semantic match	15		2	0	2	Bimodal gesture-speech vs gesture and vs speech
39	63	Kircher	2009	Figure 3B: gesture related activation increase	14		3	1	4	Facial expression & intonated spoken words, judge emotion expressed (AV > A,V; p<0.05 only)
40	64	Kreifelts	2007	Table 1 voice-face emotion	24		1	2	3	Increased effectiveness connectivity with pSTS and thalamus during AV integration of non-verbal emotional information
	65	Kreifelts	2007	Table 5 AV increase effective connectivity		pooled	2	4	6	

Downloaded from https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/igab002/6103812 by West Virginia University user on 05 February 2021

41	66	Lewis	2000	Table 1	7		2	3	5	Compare speed of tone sweeps to visual dot coherent motion: Bimodal vs unimodal
42	67	Matchin	2014	Table 1 AV > Aud only (McGurk)	20		2	7	9	McGurk audio-visual speech: AV > A only
	68	Matchin	2014	Table 2 AV > Video only		pooled	9	6	15	McGurk audio-visual speech: AV > V only
43	70	McNamara	2008	Table (BA44 and IPL)	12		2	2	4	Videos of meaningless hand gestures & synthetic tone sounds: Increases in functional connectivity with learning
49	79	Nath	2012	pg 784	14		1	0	1	McGurk effect (phonemes): congruent AV speech correlated with behavioral percept
54	87	Ogawa	2013b	Table 1 3D > 2D and surround > monaural effects	16		3	4	7	Cinematic 3D > 2D video and surround sound > monaural while watching a movie ("The Three Musketeers")
55	88	Okada	2013	Table 1 AV > A	20		5	4	9	Video of AV > A speech only
56	89	Olson	2002	Table 1A synchronized AV > static Vis-only	10		7	4	11	Whole face video & heard words: Synchronized AV vs static V
	90	Olson	2002	Table 1C synchronized AV > desynchronized AV speech		pooled	2	0	2	Whole face video & heard words: Synchronized vs desynchronized
59	94	Regenbogen	2017	Table 2A degraded > clear Multisensory vs unimodal input	29		5	6	11	Degraded > clear AV vs both visual and auditory unimodal visual real-world object-in-action recognition
60	95	Robins	2008	Table 2 (Fig 2) AV integration (AV > A and AV > V)	10		2	1	3	Face speaking sentences: angry, fearful, happy, neutral (AV > AV)
	96	Robins	2008	Table 4A (Fig 5) AV integration & emotion		pooled	1	4	5	AV faces and spoken sentences expressing fear or neutral valence: AV integration (AV > A, V conditions)
	97	Robins	2008	Table 4B emotion effects		pooled	2	0	2	AV faces and spoken sentences expressing fear or neutral valence: Emotional AV-fear > AV-neutral
	98	Robins	2008	Table 4C (Fig 5) fearful AV integration		pooled	1	5	6	AV faces and spoken sentences expressing fear or neutral valence: Fearful-only AV integration
	99	Robins	2008	Table 4D AV-only emotion		pooled	1	3	4	AV faces and spoken sentences expressing fear or neutral valence: AV-only emotion
61	100	Scheef	2009	Table 1 cartoon jump + boing	16		1	2	3	Video of cartoon person jumping & "sonification" of a tone, learn correlated pairings: AV-V and AV-A conjunction
63	104	Sekiyama	2003	Table 3 (fMRI nAV-AV)	8		1	0	1	AV speech, McGurk effect with phonemes (ba, da, ga) and noise modulation: noise-AV > AV (fMRI)
	105	Sekiyama	2003	Table 4 (PET nAV-AV)		pooled	1	3	4	AV speech, McGurk effect with phonemes (ba, da, ga) and noise modulation: noise-AV > AV (PET)
65	108	Stevenson	2009	Table 1B 2 AVtools > AVspeech	11		1	1	2	Hand tools in use video: inverse effectiveness (degraded AV tool > AV speech)
	109	Stevenson	2009	Table 1C (Fig 8) AVspeech > AVtools		pooled	1	1	2	Face & speech video: inverse effectiveness (degraded AV speech > AV tool use)
66	110	Straube	2011	Table 3A-B iconic/metaphoric speech-gestures vs speech, gestu	16		2	2	4	Integration of iconic and Metaphoric speech-gestures vs speech and gesture
67	111	Straube	2014	p939 Integration foci	16		3	0	3	Integration of iconic hand gesture-speech > unimodal speech and unimodal gesture (healthy control group)
74	123	Van der Wyk	2010	Table 2 AV interaction effects oval/circles+speech/nonspeech	16		3	3	6	Geometric shape modulate with speech (sentences)
78	128	Watson	2014a	Table 1A AV-adaptation effect (multimodal localizer)	18		0	1	1	Videos of emotional faces and voice: multisensory localizer
	129	Watson	2014a	Table 1C AV-adaptation effect, cross-modal adaptation effect		pooled	0	1	1	Videos of emotional faces and voice: crossmodal adaptation effects
79	130	Watson	2014b	Table 1 AV > baseline (Living and non-Living)	40		3	5	8	Moving objects and videos of faces with corresponding sounds: AV>baseline
	131	Watson	2014b	Table 4A integrative regions (Living and non-Living)		pooled	2	2	4	Moving objects and videos of faces with corresponding sounds: Integrative regions (AV > A, V)
	132	Watson	2014b	Table 4B integrative regions (Living and non-Living)		pooled	0	1	1	Moving objects and videos of faces with corresponding sounds: People-selective integrative region
80	133	Werner	2010	Table 1 superadditive (AV-salience effect)	21		0	3	3	Categorize movies of actions with tools or musical instruments (degraded stimuli): AV interactions both tasks
	134	Werner	2010	Table 2 AV interactions predict behavior		pooled	1	2	3	Categorize movies of actions with tools or musical instruments: AV interactions predicted by behavior
	135	Werner	2010	Table 3C superadditive AV due to task		pooled	3	0	3	Categorize movies of actions with tools or musical instruments: Subadditive AV to task

UNCORRECTED MANUSCRIPT

Downloaded from https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/igab002/6103812 by West Virginia University user on 05 February 2021

Table 13. Studies included in the Static-visual stimuli category for audio-visual interaction site meta-analyses. Results shown in Fig. 3E pink. Refer to Tables 1-2 and text for other details.

Study #	Experiment #	First author	Year	Experimental code and abbreviated task	Subjects	Multiple experiments	Left hem foci	Right hem foci	Number of foci	Brief description of experimental paradigm
26	39				405		106	89	195	
1	1	Adams	2002	Expt 1 Table 3 A+V (aud coords only)	12		5	1	6	A and V commonly showing subordinate > basic object name verification (words with pictures of environmental sounds)
6	7	Beauchamp	2004a	Fig 3J-K, Table 1 first 2 foci only	26		2	0	2	See photographs of tools, animals and hear corresponding sounds vs scrambled images and synthesized rippled sounds
8	9	Belardinelli	2004	Table 1 AV semantic congruence	13		6	6	12	Colored images of tools, animals, humans and semantically congruent vs incongruent sounds
17	24	Calvert	1999	Table 1 (Fig 1)	5		3	4	7	View image of lower face and hear numbers 1 through 10 vs unimodal conditions (AV > Photos; Auditory)
24	37	Gonzalo	2000	Table 1 AV > AVincon music & Chinese ideograms	14		1	1	2	Learn novel Kanji characters & musical chords, activity increases over time for consistent AV pairings
	39	Gonzalo	2000	Table 3 AV consistent vs Aud		pooled	1	1	2	Learn novel Kanji characters & musical chords, learn consistent (vs inconsistent) pairings versus auditory only
28	45	Hashimoto	2004	Table 1G (Fig4B, red) Learning Hangul letters to sounds	12		2	1	3	Unfamiliar Hangul letters & nonsense words, learn speech vs tone/noise pairings
31	49	Hein	2007	Fig 2B AV-artificial/non-living	18		0	1	1	B/W images of artificial objects ("fribbles") and animal vocalizations versus unimodal A, V
	50	Hein	2007	Fig 2C pSTS, pSTG, mSTG AV-cong		pooled	0	3	3	Familiar animal images & correct vocalizations (dog: woof-woof)
	52	Hein	2007	Fig 3B Foci 2, 3, 4 (blue) artificial/non-living		pooled	3	0	3	Visual "Fribbles" & backward/underwater distorted animal sounds, learn pairings (blue foci 2,3,4)
	53	Hein	2007	Fig 3C congruent living (green)		pooled	3	0	3	Familiar congruent living vs artificial AV object features and animal sounds (green foci 7, 8, 10)
32	54	Hocking	2008	pg 2444 verbal	18		2	0	2	(pSTS mask) Color photos, written names, auditory names, environmental sounds conceptually matched "amodal"
34	57	James	2003	Fig 2	12		0	1	1	Activation by visual objects ("Greebles") associated with auditory features (e.g. buzzes, screeches); (STG)
45	73	Muller	2012	Table S1 effective connectivity changes	27		4	3	7	Emotional facial expression (groaning, laughing) AV integration and gating of information
47	75	Naghavi	2007	Fig 1C	23		0	3	3	B/W pictures (animals, tools, instruments, vehicles) & their sounds: Cong vs Incong
48	76	Naghavi	2011	Fig 2A cong = incon	30		1	0	1	B/W drawings of objects (living and non) and natural sounds (barking, piano): congruent = incongruent encoding
	77	Naghavi	2011	Fig 2B congruent > incongruent		pooled	0	1	1	B/W drawings of objects (living and non) and natural sounds (barking, piano): congruent > incongruent encoding
50	80	Naumer	2008	Fig 2 Table 1A max contrast	18		8	6	14	Images of "Fribbles" & learned artificial sounds (underwater animal vocals): post training vs max contrast
	81	Naumer	2008	Fig 3 Table 1B pre-post		pooled	5	6	11	Images of "Fribbles" & learned corresponding artificial sounds: Post- vs Pre-training session
51	83	Naumer	2011	Fig 3C	10		1	0	1	Photographs of objects (living and non) and related natural sounds
52	85	Noppeny	2008	Table 3 AV congruent sounds/words	17		4	0	4	Speech sound recognition through AV priming, environmental sounds and spoken words: Congruent > incongruent
57	91	Plank	2012	pg 803 AV congruent effect	15		0	1	1	AV spatially congruent > semantically matching images of natural objects and associated sounds (right STG)
	92	Plank	2012	Table 2A spatially congruent-baseline		pooled	5	5	10	Images of natural objects and associated sounds, spatially congruent vs baseline
58	93	Raij	2000	Table 1B letters & speech sounds	9		2	3	5	Integration of visual letters and corresponding auditory phonetic expressions (MEG study) AV vs (A + V)
62	101	Schmid	2011	Table 2E A effect V (Living & non-living, pictures)	12		3	4	7	Environmental sounds & matching pictures: reduced activity by A
	102	Schmid	2011	Table 2F V competition effect A (reduced activity by a visual object)		pooled	2	2	4	Environmental sounds & matching pictures: reduced activity by V
	103	Schmid	2011	Table 2G AV crossmodal interaction x auditory attention		pooled	2	3	5	Environmental sounds and matching pictures: cross-modal interaction and auditory attention
64	106	Sestieri	2006	Table 1 (Fig 3), AV location match vs semantic	10		2	5	7	B/W images (animal, weapons) & environmental sounds: Match location > recognition
	107	Sestieri	2006	Table 2 AV semantic recognition vs localization		pooled	2	1	3	B/W pictures and environmental sounds: congruent semantic recognition > localization task
69	113	Tanabe	2005	Table 1A AV; A then V; not VA	15		10	10	20	Amorphous texture patterns & modulated white noises: Activation during learning delay period (AV)
	114	Tanabe	2005	Table 2A+2B (Fig 5a) AV and VA		pooled	5	6	11	Amorphous texture patterns & modulated white noises: changes after feedback learning (AV and VA)
	115	Tanabe	2005	Table 3A+3B (Fig 6) AV and VA; delay period		pooled	9	1	10	Amorphous texture patterns & modulated white noises: sustained activity throughout learning (AV and VA)
70	117	Taylor	2006	Fig 1A-B, Fig 1C-D (living > non-living)	15		2	0	2	Color photos (V), environmental sounds and spoken words (A): Cong AV vs Incong (living objects)
71	118	Van Atteveldt	2004	Table 1a letters & speech sounds	16		3	1	4	Familiar letters & their speech sounds: Congruent vs not and Bimodal vs Unimodal
72	119	Van Atteveldt	2007	Table 2A+B (Fig 2)	12		3	2	5	Single letters & their speech sounds (phonemes): Congruent > Incong; Passive perception, blocked and event-related
	120	Van Atteveldt	2007	Table 3 (Fig 2) passive		pooled	1	1	2	Single letters & their speech sounds (phonemes): Congruent > Incong; active perception task
73	122	Van Atteveldt	2010	Table 1B STS; specific adaptation congruent > incong	16		3	1	4	Letter and speech sound pairs (vowels, consonants): Specific adaptation effects
75	124	Von Kriegstein	2006	Fig 4B after > before voice-face	14		0	4	4	Face and object photos with voice and other sounds: Voice-Face association learning
82	137	Wolf	2014	Table 1 face cartoons + phonemes	16		1	1	2	Drawing of faces with emotional expressions: Supramodal effects with emotional valence

Table 14. Locations of significant clusters from the meta-analyses involving Dynamic-visual and Static-visual audio-visual experimental paradigms (from Tables 12 and 13), indicating major contributing studies to the ALE meta-analysis clusters, weighted centers of mass (x, y, and z) in Talairach coordinates, brain volumes (mm³), and ALE values. Single study ALE maps for (A) Dynamic-visual stimuli (corrected FWE $p < .05$) and (B) Static-visual stimuli (non-moving images) (corrected FWE $p < .05$), plus (C) contrast maps of interaction sites revealing Dynamic-visual $>$ Static-visual, and (D) Static-visual $>$ Dynamic-visual audio-visual stimuli (both uncorrected $p < .05$). The coordinates correspond to foci illustrated in Figure 3E (blue and pink hues).

Condition	Region	Major Contributing studies	x	y	z	Volume	ALE value
A. Dynamic-visual audio-visual stimuli > baseline							
1	R Posterior Superior Temporal Sulcus	9 (20,40,60,60,65,66,74,78,79)	48	-36	11	1312	0.037
2	L Posterior Superior Temporal Sulcus	6 (18,40,65,66,67,79)	-51	-49	10	928	0.035
3	L Posterior Superior Temporal Gyrus	2 (22,74)	-58	-38	12	136	0.027
4	R Superior Temporal Gyrus		58	-17	8	32	0.024
B. Static-visual audio-visual stimuli > baseline							
1	L Transverse Temporal Gyrus / Planum Temporale	5 (31,45,50,57,72)	-47	-22	7	552	0.031
2	R Superior Temporal Gyrus / Planum Temporale	2 (45,72)	53	-20	8	288	0.023
3	R Superior Temporal Gyrus		58	-29	11	120	0.021
C. Dynamic-visual > Static-visual audio-visual stimuli							
1	R Superior Temporal Gyrus/Sulcus	4 (40,60,60,66)	46	-37	12	392	2.287
D. Static-visual > Dynamic-visual audio-visual stimuli							
1	L Superior Temporal Gyrus / Planum Temporale /TTG	4 (31,50,57,72)	-46	-22	7	480	2.620
2	R Superior Temporal Gyrus (posterior)		58	-28	11	128	2.308
3	R Superior Temporal Gyrus/Sulcus		52	-20	3	64	2.254
4	R Transverse Temporal Gyrus		50	-18	8	24	1.739
5	R Transverse Temporal Gyrus		52	-18	12	8	1.863

ACKNOWLEDGMENTS

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIGMS) CoBRE grant GM103503 to the Centers for Neuroscience of West Virginia University, and affiliated WVU Summer Undergraduate Research Internships, plus an Internship for Medical Students through the National Institute of General Medical Sciences, U54GM104942-02.

Appendix A. List of all 137 experimental coordinates from the 82 studies after converting all to afni-TLRC coordinates using GingerALE software. The number of subjects are also indicated. The coordinate sets were used to derive all of the meta-analyses of the present study.

Supplementary Online materials. The volumetric data used to illustrate the regions of interest in Figure 3 are contained here in the form of nii (“nifty”) files that emerge from the GingerALE software. They are in afni-TLRC coordinate space.

REFERENCES

- Adams, R. B., & Janata, P. (2002). A comparison of neural circuits underlying auditory and visual object categorization. *Neuroimage*, *16*(2), 361-377.
- Alink, A., Singer, W., & Muckli, L. (2008). Capture of auditory motion by vision is represented by an activation shift from auditory to visual motion cortex. *J Neurosci*, *28*(11), 2690-2697.
- Amedi, A., Stern, W. M., Camprodon, J. A., Bermpohl, F., Merabet, L., Rotman, S., . . . Pascual-Leone, A. (2007). Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. *Nat Neurosci*, *10*(6), 687-689.
- Balk, M. H., Ojanen, V., Pekkola, J., Autti, T., Sams, M., & Jaaskelainen, I. P. (2010). Synchrony of audio-visual speech stimuli modulates left superior temporal sulcus. *Neuroreport*, *21*(12), 822-826. doi:10.1097/WNR.0b013e32833d138f
- Bar, M., Tootell, R. B. H., Schacter, D. L., Greve, D. N., Fischl, B., Mendola, J. D., . . . Dale, A. M. (2001). Cortical mechanisms specific to explicit visual object recognition. *Neuron*, *29*(2), 529-535.
- Barsalou, L. W. (2008). Grounded cognition. *Annu Rev Psychol*, *59*, 617-645.
- Barsalou, L. W., Kyle Simmons, W., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends Cogn Sci*, *7*(2), 84-91.
- Baumann, O., & Greenlee, M. W. (2007). Neural correlates of coherent audiovisual motion perception. *Cereb Cortex*, *17*(6), 1433-1443.
- Baumgaertner, A., Buccino, G., Lange, R., McNamara, A., & Binkofski, F. (2007). Polymodal conceptual processing of human biological actions in the left inferior frontal lobe. *Eur J Neurosci*, *25*(3), 881-889.
- Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature Neuroscience*, *7*, 1190-1192.
- Beauchamp, M. S., Lee, K. M., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*, 809-823.
- Belardinelli, M., Sestieri, C., Di Matteo, R., Delogu, F., Del Gratta, C., Ferretti, A., . . . Romani, G. (2004). Audio-visual crossmodal interactions in environmental perception: an fMRI investigation. *Cogn Process*, *5*, 167-174.
- Belin, P., & Zatorre, R. (2000). 'What', 'where' and 'how' in auditory cortex. *Nature Neuroscience*, *3*(10), 965-966.
- Biau, E., Moris Fernandez, L., Holle, H., Avila, C., & Soto-Faraco, S. (2016). Hand gestures as visual prosody: BOLD responses to audio-visual alignment are modulated by the communicative nature of the stimuli. *Neuroimage*, *132*, 129-137. doi:10.1016/j.neuroimage.2016.02.018
- Bischoff, M., Walter, B., Blecker, C. R., Morgen, K., Vaitl, D., & Sammer, G. (2007). Utilizing the ventriloquism-effect to investigate audio-visual binding. *Neuropsychologia*, *45*(3), 578-586. doi:10.1016/j.neuropsychologia.2006.03.008

- Blank, H., & von Kriegstein, K. (2013). Mechanisms of enhancing visual-speech recognition by prior auditory information. *Neuroimage*, *65*, 109-118. doi:10.1016/j.neuroimage.2012.09.047
- Bonath, B., Noesselt, T., Krauel, K., Tyll, S., Tempelmann, C., & Hillyard, S. A. (2014). Audio-visual synchrony modulates the ventriloquist illusion and its neural/spatial representation in the auditory cortex. *Neuroimage*, *98*, 425-434. doi:10.1016/j.neuroimage.2014.04.077
- Bonath, B., Tyll, S., Budinger, E., Krauel, K., Hopf, J. M., & Noesselt, T. (2013). Task-demands and audio-visual stimulus configurations modulate neural activity in the human thalamus. *Neuroimage*, *66*, 110-118. doi:10.1016/j.neuroimage.2012.10.018
- Brefczynski-Lewis, J. A., & Lewis, J. W. (2017). Auditory object perception: A neurobiological model and prospective review. *Neuropsychologia*, *105*, 223-242. doi:10.1016/j.neuropsychologia.2017.04.034
- Bulkin, D. A., & Groh, J. M. (2006). Seeing sounds: visual and auditory interactions in the brain. *Curr Opin Neurobiol*, *16*(4), 415-419.
- Burton, H., Snyder, A. Z., & Raichle, M. E. (2004). Default brain functionality in blind people. *Proc Natl Acad Sci U S A*, *101*(43), 15500-15505.
- Bushara, K. O., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *J Neurosci*, *21*(1), 300-304.
- Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., & Hallett, M. (2003). Neural correlates of cross-modal binding. *Nat Neurosci*, *6*(2), 190-195.
- Cacioppo, S. (2013). Selective decision-making deficit in love following damage to the anterior insula. *Front Hum Neurosci*, *7*, 15-19. doi:10.3389/fnhum.2013.00099
- Callan, D. E., Jones, J. A., & Callan, A. (2014). Multisensory and modality specific processing of visual speech in different regions of the premotor cortex. *Front Psychol*, *5*, 389. doi:10.3389/fpsyg.2014.00389
- Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport*, *10*(12), 2619-2623.
- Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *J Cogn Neurosci*, *15*(1), 57-70.
- Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*, 649-657.
- Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage*, *14*(2), 427-438.
- Calvert, G. A., & Lewis, J. W. (2004). Hemodynamic studies of audio-visual interactions. In G. A. Calvert, C. Spence, & B. Stein (Eds.), *Handbook of multisensory processing* (pp. 483-502). Cambridge, Massachusetts: MIT Press.
- Capek, C. M., Macsweeney, M., Woll, B., Waters, D., McGuire, P. K., David, A. S., . . . Campbell, R. (2008). Cortical circuits for silent speechreading in deaf and hearing people. *Neuropsychologia*, *46*(5), 1233-1241.
- Capek, C. M., Woll, B., MacSweeney, M., Waters, D., McGuire, P. K., David, A. S., . . . Campbell, R. (2010). Superior temporal activation as a function of linguistic

- knowledge: insights from deaf native signers who speechread. *Brain Lang*, 112(2), 129-134. doi:10.1016/j.bandl.2009.10.004
- Caramazza, A., & Mahon, B. Z. (2003). The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends Cogn Sci*, 7(8), 354-361.
- Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain the animate-inanimate distinction. *J Cogn Neurosci*, 10(1), 1-34.
- Cascella, N. G., Gerner, G. J., Fieldstone, S. C., Sawa, A., & Schretlen, D. J. (2011). The insula-claustrum region and delusions in schizophrenia. *Schizophr Res*, 133(1-3), 77-81. doi:10.1016/j.schres.2011.08.004
- Cecere, R., Gross, J., & Thut, G. (2016). Behavioural evidence for separate mechanisms of audiovisual temporal binding as a function of leading sensory modality. *Eur J Neurosci*, 43(12), 1561-1568. doi:10.1111/ejn.13242
- Clarke, S., Thiran, A. B., Maeder, P., Adriani, M., Vernet, O., Regli, L., . . . Thiran, J.-P. (2002). What and where in human audition: selective deficits following focal hemispheric lesions. *Experimental Brain Research*, 147, 8-15.
- Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29, 162-173.
- Craig, A. D. (2009). How do you feel--now? The anterior insula and human awareness. *Nat Rev Neurosci*, 10(1), 59-70.
- Craig, A. D. (2010). The sentient self. *Brain Struct Funct*, 214(5-6), 563-577. doi:10.1007/s00429-010-0248-y
- Crick, F. C., & Koch, C. (2005). What is the function of the claustrum? *Philos Trans R Soc Lond B Biol Sci*, 360(1458), 1271-1279.
- Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nat Neurosci*, 7(2), 189-195. doi:10.1038/nrn1176
- Dalgleish, T. (2004). The emotional brain. *Nat Rev Neurosci*, 5(7), 583-589. doi:10.1038/nrn1432
- Damasio, A. (2001). Fundamental feelings. *Nature*, 413(6858), 781. doi:10.1038/35101669
- Damasio, A. R. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural Comput.*, 1, 123-132.
- Damasio, A. R. (1989). Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33(1-2), 25-62. doi:10.1016/0010-0277(89)90005-x
- Damasio, A. R. (1999). How the brain creates the mind. *Sci Am*, 281(6), 112-117.
- Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., & Damasio, R. D. (1996). A neural basis for lexical retrieval. *Nature*, 380, 499-505.
- de Haas, B., Schwarzkopf, D. S., Uner, M., & Rees, G. (2013). Auditory modulation of visual stimulus encoding in human retinotopic cortex. *Neuroimage*, 70, 258-267. doi:10.1016/j.neuroimage.2012.12.061
- Dick, F., Saygin, A. P., Galati, G., Pitzalis, S., Bentrovato, S., D'Amico, S., . . . Pizzamiglio, L. (2007). What is involved and what is necessary for complex linguistic and nonlinguistic auditory processing: evidence from functional magnetic resonance imaging and lesion data. *J Cogn Neurosci*, 19(5), 799-816.

- Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*: Harvard University Press.
- Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., & Fox, P. T. (2012). Activation likelihood estimation meta-analysis revisited. *Neuroimage*, *59*(3), 2349-2361. doi:10.1016/j.neuroimage.2011.09.017
- Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., & Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum Brain Mapp*, *30*(9), 2907-2926. doi:10.1002/hbm.20718
- Eickhoff, S. B., Nichols, T. E., Laird, A. R., Hoffstaedter, F., Amunts, K., Fox, P. T., . . . Eickhoff, C. R. (2016). Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *Neuroimage*, *137*, 70-85. doi:10.1016/j.neuroimage.2016.04.072
- Engel, L. R., Frum, C., Puce, A., Walker, N. A., & Lewis, J. W. (2009). Different categories of living and non-living sound-sources activate distinct cortical networks. *Neuroimage*, *47*(4), 1778-1791.
- Erickson, L. C., Zielinski, B. A., Zielinski, J. E., Liu, G., Turkeltaub, P. E., Leaver, A. M., & Rauschecker, J. P. (2014). Distinct cortical locations for integration of audiovisual speech and the McGurk effect. *Front Psychol*, *5*, 534. doi:10.3389/fpsyg.2014.00534
- Ethofer, T., Bretschner, J., Wiethoff, S., Bisch, J., Schlipf, S., Wildgruber, D., & Kreifelts, B. (2013). Functional responses and structural connections of cortical areas for processing faces and voices in the superior temporal sulcus. *Neuroimage*, *76*, 45-56. doi:10.1016/j.neuroimage.2013.02.064
- Galati, G., Committeri, G., Spitoni, G., Aprile, T., Di Russo, F., Pitzalis, S., & Pizzamiglio, L. (2008). A selective representation of the meaning of actions in the auditory mirror system. *Neuroimage*, *40*(3), 1274-1286. doi:10.1016/j.neuroimage.2007.12.044
- Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nat Neurosci*, *2*(6), 568-573.
- Gazzola, V., Aziz-Zadeh, L., & Keysers, C. (2006). Empathy and the somatotopic auditory mirror system in humans. *Curr Biol*, *16*(18), 1824-1829.
- Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends Cogn Sci*, *10*(6), 278-285. doi:10.1016/j.tics.2006.04.008
- Goll, J. C., Crutch, S. J., & Warren, J. D. (2011). Central auditory disorders: toward a neuropsychology of auditory objects. *Curr Opin Neurol*, *23*(6), 617-627.
- Gonzalo, D., Shallice, T., & Dolan, R. (2000). Time-dependent changes in learning audiovisual associations: a single-trial fMRI study. *Neuroimage*, *11*(3), 243-255.
- Goodale, M. A., Meenan, J. P., Bulthoff, H. H., Nicolle, D. A., Murphy, K. J., & Racicot, C. I. (1994). Separate neural pathways for the visual analysis of object shape in perception and prehension. *Curr Biol*, *4*(7), 604-610.
- Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*, 20-25.

- Green, A., Straube, B., Weis, S., Jansen, A., Willmes, K., Konrad, K., & Kircher, T. (2009). Neural integration of iconic and unrelated coverbal gestures: a functional MRI study. *Hum Brain Mapp*, *30*(10), 3309-3324. doi:10.1002/hbm.20753
- Hagan, C. C., Woods, W., Johnson, S., Green, G. G., & Young, A. W. (2013). Involvement of right STS in audio-visual integration for affective speech demonstrated using MEG. *PLoS One*, *8*(8), e70648. doi:10.1371/journal.pone.0070648
- Happe, F., & Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *J Autism Dev Disord*, *36*(1), 5-25. doi:10.1007/s10803-005-0039-0
- Hasegawa, T., Matsuki, K., Ueno, T., Maeda, Y., Matsue, Y., Konishi, Y., & Sadato, N. (2004). Learned audio-visual cross-modal associations in observed piano playing activate the left planum temporale. An fMRI study. *Brain Res Cogn Brain Res*, *20*(3), 510-518. doi:10.1016/j.cogbrainres.2004.04.005
- Hashimoto, R., & Sakai, K. L. (2004). Learning letters in adulthood: direct visualization of cortical plasticity for forming a new link between orthography and phonology. *Neuron*, *42*(2), 311-322.
- He, Y., Gebhardt, H., Steines, M., Sammer, G., Kircher, T., Nagels, A., & Straube, B. (2015). The EEG and fMRI signatures of neural integration: An investigation of meaningful gestures and corresponding speech. *Neuropsychologia*, *72*, 27-42. doi:10.1016/j.neuropsychologia.2015.04.018
- He, Y., Steines, M., Sommer, J., Gebhardt, H., Nagels, A., Sammer, G., . . . Straube, B. (2018). Spatial-temporal dynamics of gesture-speech integration: a simultaneous EEG-fMRI study. *Brain Struct Funct*, *223*(7), 3073-3089. doi:10.1007/s00429-018-1674-5
- Hein, G., Doehrmann, O., Muller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *J Neurosci*, *27*(30), 7881-7887.
- Hewes, G. W. (1973). Primate Communication and the Gestural Origin of Language. *Current Anthropology*, *14*, 5-24. doi:Doi 10.1086/204019
- Hocking, J., & Price, C. J. (2008). The role of the posterior superior temporal sulcus in audiovisual processing. *Cereb Cortex*, *18*(10), 2439-2449.
- Hove, M. J., Fairhurst, M. T., Kotz, S. A., & Keller, P. E. (2013). Synchronizing with auditory and visual rhythms: an fMRI assessment of modality differences and modality appropriateness. *Neuroimage*, *67*, 313-321. doi:10.1016/j.neuroimage.2012.11.032
- James, T. W., & Gauthier, I. (2003). Auditory and action semantic features activate sensory-specific perceptual brain regions. *Curr Biol*, *13*(20), 1792-1796.
- James, T. W., VanDerKlok, R. M., Stevenson, R. A., & James, K. H. (2011). Multisensory perception of action in posterior temporal and parietal cortices. *Neuropsychologia*, *49*(1), 108-114. doi:10.1016/j.neuropsychologia.2010.10.030
- Jellema, T., & Perrett, D. I. (2006). Neural representations of perceived bodily actions using a categorical frame of reference. *Neuropsychologia*, *44*(9), 1535-1546.
- Jessen, S., & Kotz, S. A. (2015). Affect differentially modulates brain activation in uni- and multisensory body-voice perception. *Neuropsychologia*, *66*, 134-143. doi:10.1016/j.neuropsychologia.2014.10.038

- Jola, C., McAleer, P., Grosbras, M. H., Love, S. A., Morison, G., & Pollick, F. E. (2013). Uni- and multisensory brain areas are synchronised across spectators when watching unedited dance recordings. *Iperception*, 4(4), 265-284. doi:10.1068/i0536
- Jolliffe, T., & Baron-Cohen, S. (2000). Linguistic processing in high-functioning adults with autism or Asperger's syndrome. Is global coherence impaired? *Psychol Med*, 30(5), 1169-1187.
- Kaas, J. H., & Hackett, T. A. (1999). 'What' and 'where' processing in auditory cortex. *Nat Neurosci*, 2(12), 1045-1047.
- Kellenbach, M. L., Brett, M., & Patterson, K. (2003). Actions speak louder than functions: The importance of manipulability and action in tool representation. *Journal of Cognitive Neuroscience*, 15(1), 30-46.
- Kiefer, M., Sim, E. J., Herrnberger, B., Grothe, J., & Hoenig, K. (2008). The sound of concepts: four markers for a link between auditory and conceptual brain systems. *J Neurosci*, 28(47), 12224-12230.
- Kim, H., Hahm, J., Lee, H., Kang, E., Kang, H., & Lee, D. S. (2015). Brain networks engaged in audiovisual integration during speech perception revealed by persistent homology-based network filtration. *Brain Connect*, 5(4), 245-258. doi:10.1089/brain.2013.0218
- Kircher, T., Straube, B., Leube, D., Weis, S., Sachs, O., Willmes, K., . . . Green, A. (2009). Neural interaction of speech and gesture: differential activations of metaphoric co-verbal gestures. *Neuropsychologia*, 47(1), 169-179. doi:10.1016/j.neuropsychologia.2008.08.009
- Kouijzer, M. E. J., de Moor, J. M. H., Gerrits, B. J. L., Congedo, M., & van Schie, H. T. (2009). Neurofeedback improves executive functioning in children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 3(1), 145-162. doi:10.1016/j.rasd.2008.05.001
- Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. *Neuroimage*, 37(4), 1445-1456.
- Lahav, A., Saltzman, E., & Schlaug, G. (2007). Action representation of sound: audiomotor recognition network while listening to newly acquired actions. *J Neurosci*, 27(2), 308-314. doi:10.1523/JNEUROSCI.4822-06.2007
- Laird, A. R., Eickhoff, S. B., Kurth, F., Fox, P. M., Uecker, A. M., Turner, J. A., . . . Fox, P. T. (2009). ALE Meta-Analysis Workflows Via the Brainmap Database: Progress Towards A Probabilistic Functional Brain Atlas. *Front Neuroinform*, 3, 23. doi:10.3389/neuro.11.023.2009
- Laird, N., Fitzmaurice, G., & Ding, X. (2010). Comments on 'Empirical vs natural weighting in random effects meta-analysis'. *Stat Med*, 29(12), 1266-1267; discussion 1272-1281. doi:10.1002/sim.3657
- Lamm, C., & Singer, T. (2010). The role of anterior insular cortex in social emotions. *Brain Struct Funct*, 214(5-6), 579-591. doi:10.1007/s00429-010-0251-3
- Languis, M. L., & Miller, D. C. (1992). Luria's Theory of Brain Functioning: A Model for Research in Cognitive Psychophysiology. *Educational Psychologist*, 27(4), 493-511. doi:10.1207/s15326985ep2704_6

- Lewis, J. W. (2010). Audio-visual perception of everyday natural objects—hemodynamic studies in humans. In M. J. Naumer & J. Kaiser (Eds.), *Multisensory object perception in the primate brain* (pp. 155-190). Oxford University Press: Springer.
- Lewis, J. W., Beauchamp, M. S., & DeYoe, E. A. (2000). A comparison of visual and auditory motion processing in human cerebral cortex. *Cerebral Cortex*, *10*(9), 873-888.
- Lewis, J. W., Frum, C., Brefczynski-Lewis, J. A., Talkington, W. J., Walker, N. A., Rapuano, K. M., & Kovach, A. L. (2011). Cortical Network Differences in the Sighted Versus Early Blind for Recognition of Human-Produced Action Sounds. *Human Brain Mapping*, *32*(12), 2241-2255. doi:10.1002/hbm.21185
- Lewis, J. W., Phinney, R. E., Brefczynski-Lewis, J. A., & DeYoe, E. A. (2006). Lefties get it "right" when hearing tool sounds. *J Cogn Neurosci*, *18*(8), 1314-1330. doi:10.1162/jocn.2006.18.8.1314
- Lewis, J. W., Silberman, M. J., Donai, J. J., Frum, C. A., & Brefczynski-Lewis, J. A. (2018). Hearing and orally mimicking different acoustic-semantic categories of natural sound engage distinct left hemisphere cortical regions. *Brain Lang*, *183*, 64-78. doi:10.1016/j.bandl.2018.05.002
- Lewis, J. W., Talkington, W. J., Puce, A., Engel, L. R., & Frum, C. (2011). Cortical networks representing object categories and high-level attributes of familiar real-world action sounds. *J Cogn Neurosci*, *23*(8), 2079-2101. doi:10.1162/jocn.2010.21570
- Lewis, J. W., Talkington, W. J., Tallaksen, K. C., & Frum, C. A. (2012). Auditory object salience: human cortical processing of non-biological action sounds and their acoustic signal attributes. *Front Syst Neurosci*, *6*(27)(27), 1-16.
- Lewis, J. W., Wightman, F. L., Brefczynski, J. A., Phinney, R. E., Binder, J. R., & DeYoe, E. A. (2004). Human brain regions involved in recognizing environmental sounds. *Cerebral Cortex*, *14*, 1008-1021.
- Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin*, *126*, 281-308.
- Lissauer, H. (1890/1988). A case of visual agnosia with a contribution to theory. *Cognitive Neuropsychology*, *5*, 157-192.
- Lutz, A., Brefczynski-Lewis, J., Johnstone, T., & Davidson, R. J. (2008). Regulation of the neural circuitry of emotion by compassion meditation: effects of meditative expertise. *PLoS One*, *3*(3), e1897. doi:10.1371/journal.pone.0001897
- Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M., & Caramazza, A. (2009). Category-specific organization in the human brain does not require visual experience. *Neuron*, *63*(3), 397-405.
- Mahon, B. Z., & Caramazza, A. (2005). The orchestration of the sensory-motor systems: Clues from neuropsychology. *Cognitive Neuropsychology*, *22*, 480-494.
- Marco, E. J., Hinkley, L. B., Hill, S. S., & Nagarajan, S. S. (2011). Sensory processing in autism: a review of neurophysiologic findings. *Pediatr Res*, *69*(5 Pt 2), 48R-54R. doi:10.1203/PDR.0b013e3182130c54
- Martin, A. (2007). The representation of object concepts in the brain. *Annu Rev Psychol*, *58*, 25-45.
- Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature*, *379*(6566), 649-652.

- Matchin, W., Groulx, K., & Hickok, G. (2014). Audiovisual speech integration does not rely on the motor system: evidence from articulatory suppression, the McGurk effect, and fMRI. *J Cogn Neurosci*, *26*(3), 606-620. doi:10.1162/jocn_a_00515
- McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nat Rev Neurosci*, *4*(4), 310-322.
- McNamara, A., Buccino, G., Menz, M. M., Glascher, J., Wolbers, T., Baumgartner, A., & Binkofski, F. (2008). Neural dynamics of learning sound-action associations. *PLoS One*, *3*(12), e3845.
- Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Struct Funct*, *214*(5-6), 655-667. doi:10.1007/s00429-010-0262-0
- Meyer, M., Baumann, S., Marchina, S., & Jancke, L. (2007). Hemodynamic responses in human multisensory and auditory association cortex to purely visual stimulation. *BMC Neurosci*, *8*, 14.
- Miller, E. K., Nieder, A., Freedman, D. J., & Wallis, J. D. (2003). Neural correlates of categories and concepts. *Curr Opin Neurobiol*, *13*(2), 198-203.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med*, *151*(4), 264-269, W264.
- Molenberghs, P., Cunnington, R., & Mattingley, J. B. (2012). Brain regions with mirror properties: a meta-analysis of 125 human fMRI studies. *Neurosci Biobehav Rev*, *36*(1), 341-349. doi:10.1016/j.neubiorev.2011.07.004
- Muller, V. I., Cieslik, E. C., Laird, A. R., Fox, P. T., Radua, J., Mataix-Cols, D., . . . Eickhoff, S. B. (2018). Ten simple rules for neuroimaging meta-analysis. *Neurosci Biobehav Rev*, *84*, 151-161. doi:10.1016/j.neubiorev.2017.11.012
- Muller, V. I., Cieslik, E. C., Turetsky, B. I., & Eickhoff, S. B. (2012). Crossmodal interactions in audiovisual emotion processing. *Neuroimage*, *60*(1), 553-561. doi:10.1016/j.neuroimage.2011.12.007
- Murase, M., Saito, D. N., Kochiyama, T., Tanabe, H. C., Tanaka, S., Harada, T., . . . Sadato, N. (2008). Cross-modal integration during vowel identification in audiovisual speech: a functional magnetic resonance imaging study. *Neurosci Lett*, *434*(1), 71-76.
- Mutschler, I., Wieckhorst, B., Kowalevski, S., Derix, J., Wentlandt, J., Schulze-Bonhage, A., & Ball, T. (2009). Functional organization of the human anterior insular cortex. *Neurosci Lett*, *457*(2), 66-70.
- Naghavi, H. R., Eriksson, J., Larsson, A., & Nyberg, L. (2007). The claustrum/insula region integrates conceptually related sounds and pictures. *Neurosci Lett*, *422*(1), 77-80.
- Naghavi, H. R., Eriksson, J., Larsson, A., & Nyberg, L. (2011). Cortical regions underlying successful encoding of semantically congruent and incongruent associations between common auditory and visual objects. *Neurosci Lett*, *505*(2), 191-195. doi:10.1016/j.neulet.2011.10.022
- Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage*, *59*(1), 781-787. doi:10.1016/j.neuroimage.2011.07.024

- Naumer, M. J., Doehrmann, O., Muller, N. G., Muckli, L., Kaiser, J., & Hein, G. (2008). Cortical Plasticity of Audio-Visual Object Representations. *Cereb Cortex*, *19*(7), 1641-1653.
- Naumer, M. J., van den Bosch, J. J., Wibrals, M., Kohler, A., Singer, W., Kaiser, J., . . . Muckli, L. (2011). Investigating human audio-visual object perception with a combination of hypothesis-generating and hypothesis-testing fMRI analysis tools. *Exp Brain Res*, *213*(2-3), 309-320. doi:10.1007/s00221-011-2669-0
- Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cereb Cortex*, *18*(3), 598-609.
- Ogawa, A., Bordier, C., & Macaluso, E. (2013). Audio-visual perception of 3D cinematography: an fMRI study using condition-based and computation-based analyses. *PLoS One*, *8*(10), e76003. doi:10.1371/journal.pone.0076003
- Ogawa, A., & Macaluso, E. (2013). Audio-visual interactions for motion perception in depth modulate activity in visual area V3A. *Neuroimage*, *71*, 158-167. doi:10.1016/j.neuroimage.2013.01.012
- Okada, K., Venezia, J. H., Matchin, W., Saberi, K., & Hickok, G. (2013). An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex. *PLoS One*, *8*(6), e68959. doi:10.1371/journal.pone.0068959
- Olson, I. R., Gatenby, J. C., & Gore, J. C. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Brain Res Cogn Brain Res*, *14*(1), 129-138.
- Pascual-Leone, A., & Hamilton, R. (2001). The metamodal organization of the brain. *Prog Brain Res*, *134*, 427-445.
- Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat Rev Neurosci*, *8*(12), 976-987.
- Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *J Cogn Neurosci*, *16*(10), 1706-1716.
- Pfeiffer, B., Clark, G. F., & Arbesman, M. (2018). Effectiveness of Cognitive and Occupation-Based Interventions for Children With Challenges in Sensory Processing and Integration: A Systematic Review. *Am J Occup Ther*, *72*(1), 7201190020p7201190021-7201190020p7201190029. doi:10.5014/ajot.2018.028233
- Pfeiffer, B. A., Koenig, K., Kinnealey, M., Sheppard, M., & Henderson, L. (2011). Effectiveness of sensory integration interventions in children with autism spectrum disorders: a pilot study. *Am J Occup Ther*, *65*(1), 76-85.
- Pietrini, P., Furey, M. L., Ricciardi, E., Gobbin, M. I., Wu, W. H., Cohen, L., . . . Haxby, J. V. (2004). Beyond sensory images: Object-based representation in the human ventral pathway. *Proc Natl Acad Sci U S A*, *101*(15), 5658-5663.
- Plank, T., Rosengarth, K., Song, W., Ellermeier, W., & Greenlee, M. W. (2012). Neural correlates of audio-visual object recognition: effects of implicit spatial congruency. *Hum Brain Mapp*, *33*(4), 797-811. doi:10.1002/hbm.21254

- Powers, A. R., 3rd, Hillock, A. R., & Wallace, M. T. (2009). Perceptual training narrows the temporal window of multisensory binding. *J Neurosci*, *29*(39), 12265-12274. doi:10.1523/JNEUROSCI.3501-09.2009
- Pulvermuller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends Cogn Sci*, *17*(9), 458-470. doi:10.1016/j.tics.2013.06.004
- Pulvermuller, F. (2018). Neural reuse of action perception circuits for language, concepts and communication. *Prog Neurobiol*. doi:10.1016/j.pneurobio.2017.07.001
- Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron*, *28*(2), 617-625.
- Ramot, M., Kimmich, S., Gonzalez-Castillo, J., Roopchansingh, V., Popal, H., White, E., . . . Martin, A. (2017). Direct modulation of aberrant brain network connectivity through real-time NeuroFeedback. *Elife*, *6*. doi:10.7554/eLife.28974
- Rauschecker, J. P. (1998). Parallel processing in the auditory cortex of primates. *Audiol Neurootol*, *3*(2-3), 86-103.
- Rauschecker, J. P., & Scott, S. K. (2015). Pathways and streams in the auditory cortex. In G. S. Hickok & S. L. Small (Eds.), *In: Neurobiology of Language* (pp. 287-298): Elsevier.
- Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences USA*, *97*(22), 11800-11806.
- Regenbogen, C., Seubert, J., Johansson, E., Finkelmeyer, A., Andersson, P., & Lundstrom, J. N. (2018). The intraparietal sulcus governs multisensory integration of audiovisual information based on task difficulty. *Hum Brain Mapp*, *39*(3), 1313-1326. doi:10.1002/hbm.23918
- Rilling, J. K. (2008). Neuroscientific approaches and applications within anthropology. *Am J Phys Anthropol*, *Suppl 47*, 2-32.
- Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends Neurosci*, *21*(5), 188-194.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annu Rev Neurosci*, *27*, 169-192.
- Roa Romero, Y., Keil, J., Balz, J., Gallinat, J., & Senkowski, D. (2016). Reduced frontal theta oscillations indicate altered crossmodal prediction error processing in schizophrenia. *J Neurophysiol*, *116*(3), 1396-1407. doi:10.1152/jn.00096.2016
- Robertson, C. E., & Baron-Cohen, S. (2017). Sensory perception in autism. *Nat Rev Neurosci*, *18*(11), 671-684. doi:10.1038/nrn.2017.112
- Robins, D. L., Hunyadi, E., & Schultz, R. T. (2009). Superior temporal activation in response to dynamic audio-visual emotional cues. *Brain Cogn*, *69*(2), 269-278.
- Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*, 328-350.
- Saygin, A. P., Leech, R., & Dick, F. (2010). Nonverbal auditory agnosia with lesion to Wernicke's area. *Neuropsychologia*, *48*(1), 107-113.
- Scheef, L., Boecker, H., Daamen, M., Fehse, U., Landsberg, M. W., Granath, D. O., . . . Effenberg, A. O. (2009). Multimodal motion processing in area V5/MT: evidence from an artificial class of audio-visual events. *Brain Res*, *1252*, 94-104.

- Schmid, C., Buchel, C., & Rose, M. (2011). The neural basis of visual dominance in the context of audio-visual object processing. *Neuroimage*, *55*(1), 304-311. doi:10.1016/j.neuroimage.2010.11.051
- Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neurosci Res*, *47*(3), 277-287.
- Sestieri, C., Di Matteo, R., Ferretti, A., Del Gratta, C., Caulo, M., Tartaro, A., . . . Romani, G. L. (2006). "What" versus "where" in the audiovisual domain: an fMRI study. *Neuroimage*, *33*(2), 672-680.
- Shura, R. D., Hurley, R. A., & Taber, K. H. (2014). Insular cortex: structural and functional neuroanatomy. *J Neuropsychiatry Clin Neurosci*, *26*(4), 276-282. doi:10.1176/appi.neuropsych.260401
- Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*(5661), 1157-1162. doi:10.1126/science.1093535
- Sporns, O., Honey, C. J., & Kotter, R. (2007). Identification and classification of hubs in brain networks. *PLoS One*, *2*(10), e1049. doi:10.1371/journal.pone.0001049
- Stein, B. E., & Meredith, M. A. (1990). Multisensory integration. *Annals of the New York Academy of Sciences*, *608*(608), 51-70.
- Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.
- Stein, B. E., & Wallace, M. T. (1996). Comparisons of cross-modality integration in midbrain and cortex. *Progress in Brain Research*, *112*, 289-299.
- Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage*, *44*(3), 1210-1223.
- Straube, B., Green, A., Bromberger, B., & Kircher, T. (2011). The differentiation of iconic and metaphoric gestures: common and unique integration processes. *Hum Brain Mapp*, *32*(4), 520-533. doi:10.1002/hbm.21041
- Straube, B., Green, A., Sass, K., & Kircher, T. (2014). Superior temporal sulcus disconnectivity during processing of metaphoric gestures in schizophrenia. *Schizophr Bull*, *40*(4), 936-944. doi:10.1093/schbul/sbt110
- Szycik, G. R., Jansma, H., & Munte, T. F. (2009). Audiovisual integration during speech comprehension: an fMRI study comparing ROI-based and whole brain analyses. *Hum Brain Mapp*, *30*(7), 1990-1999. doi:10.1002/hbm.20640
- Tanabe, H. C., Honda, M., & Sadato, N. (2005). Functionally segregated neural substrates for arbitrary audiovisual paired-association learning. *J Neurosci*, *25*(27), 6409-6418.
- Taylor, K. I., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2006). Binding crossmodal object features in perirhinal cortex. *Proc Natl Acad Sci U S A*, *103*(21), 8239-8244.
- Taylor, K. I., Stamatakis, E. A., & Tyler, L. K. (2009). Crossmodal integration of object features: Voxel-based correlations in brain-damaged patients. *Brain*, *132*, 671-683.
- Tench, C. R., Tanasescu, R., Auer, D. P., Cottam, W. J., & Constantinescu, C. S. (2014). Coordinate based meta-analysis of functional neuroimaging data using activation

- likelihood estimation; full width half max and group comparisons. *PLoS One*, 9(9), e106735. doi:10.1371/journal.pone.0106735
- Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., . . . Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *J Cogn Neurosci*, 17(2), 273-281.
- Tomasello, R., Garagnani, M., Wennekers, T., & Pulvermuller, F. (2017). Brain connections of words, perceptions and actions: A neurobiological model of spatio-temporal semantic activation in the human cortex. *Neuropsychologia*, 98, 111-129. doi:10.1016/j.neuropsychologia.2016.07.004
- Tranel, D., Damasio, H., & Damasio, A. R. (1997). A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia*, 35(10), 1319-1327.
- Tranel, D., Damasio, H., Eichhorn, G. R., Grabowski, T. J., Ponto, L. L. B., & Hichwa, R. D. (2003). Neural correlates of naming animals from their characteristic sounds. *Neuropsychologia*, 41, 847-854.
- Trumpp, N. M., Kliese, D., Hoenig, K., Haarmeier, T., & Kiefer, M. (2013). Losing the sound of concepts: damage to auditory association cortex impairs the processing of sound-related concepts. *Cortex*, 49(2), 474-486. doi:10.1016/j.cortex.2012.02.002
- Turkeltaub, P. E., Eden, G. F., Jones, K. M., & Zeffiro, T. A. (2002). Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage*, 16(3 Pt 1), 765-780.
- Turkeltaub, P. E., Eickhoff, S. B., Laird, A. R., Fox, M., Wiener, M., & Fox, P. (2012). Minimizing within-experiment and within-group effects in Activation Likelihood Estimation meta-analyses. *Hum Brain Mapp*, 33(1), 1-13. doi:10.1002/hbm.21186
- Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends Cogn Sci*, 5(6), 244-252.
- Tyler, L. K., Stamatakis, E. A., Bright, P., Acres, K., Abdallah, S., Rodd, J. M., & Moss, H. E. (2004). Processing objects at different levels of specificity. *J Cogn Neurosci*, 16(3), 351-362.
- Ungerleider, L. G., & Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Curr Opin Neurobiol*, 4(2), 157-165.
- Ungerleider, L. G., Mishkin, M., Goodale, M. A., & Mansfield, R. J. W. (1982). Two cortical visual systems. In D. J. Ingle (Ed.), *Analysis of Visual Behavior* (pp. 549-586). Cambridge, MA: MIT Press.
- van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron*, 43(2), 271-282.
- van Atteveldt, N. M., Blau, V. C., Blomert, L., & Goebel, R. (2010). fMR-adaptation indicates selectivity to audiovisual content congruency in distributed clusters in human superior temporal cortex. *BMC Neurosci*, 11, 11. doi:10.1186/1471-2202-11-11
- van Atteveldt, N. M., Formisano, E., Goebel, R., & Blomert, L. (2007). Top-down task effects overrule automatic multisensory responses to letter-sound pairs in auditory association cortex. *Neuroimage*, 36(4), 1345-1360.
- van den Heuvel, M. P., & Sporns, O. (2013). Network hubs in the human brain. *Trends Cogn Sci*, 17(12), 683-696. doi:10.1016/j.tics.2013.09.012

- Van der Stoep, N., Van der Stigchel, S., Van Engelen, R. C., Biesbroek, J. M., & Nijboer, T. C. W. (2019). Impairments in Multisensory Integration after Stroke. *J Cogn Neurosci*, *31*(6), 885-899. doi:10.1162/jocn_a_01389
- Van Essen, D. C. (2005). A Population-Average, Landmark- and Surface-based (PALS) atlas of human cerebral cortex. *Neuroimage*, *28*(3), 635-662.
- Van Essen, D. C., Drury, H. A., Dickson, J., Harwell, J., Hanlon, D., & Anderson, C. H. (2001). An integrated software suite for surface-based analyses of cerebral cortex. *Journal of the American Medical Informatics Association*, *8*(5), 443-459.
- Vander Wyk, B. C., Ramsay, G. J., Hudac, C. M., Jones, W., Lin, D., Klin, A., . . . Pelphrey, K. A. (2010). Cortical integration of audio-visual speech and non-speech stimuli. *Brain Cogn*, *74*(2), 97-106. doi:10.1016/j.bandc.2010.07.002
- Vanes, L. D., White, T. P., Wigton, R. L., Joyce, D., Collier, T., & Shergill, S. S. (2016). Reduced susceptibility to the sound-induced flash fusion illusion in schizophrenia. *Psychiatry Res*, *245*, 58-65. doi:10.1016/j.psychres.2016.08.016
- von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biol*, *4*(10), e326.
- Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, Massachusetts: Harvard University Press.
- Watkins, S., Shams, L., Josephs, O., & Rees, G. (2007). Activity in human V1 follows multisensory perception. *Neuroimage*, *37*(2), 572-578.
- Watkins, S., Shams, L., Tanaka, S., Haynes, J. D., & Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. *Neuroimage*, *31*(3), 1247-1256.
- Watson, R., Latinus, M., Charest, I., Crabbe, F., & Belin, P. (2014). People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex*, *50*, 125-136. doi:10.1016/j.cortex.2013.07.011
- Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal adaptation in right posterior superior temporal sulcus during face-voice emotional integration. *J Neurosci*, *34*(20), 6813-6821. doi:10.1523/JNEUROSCI.4478-13.2014
- Webster, P. J., Frum, C., Kurowski-Burt, A., Bauer, C. E., Wen, S., Ramadan, J. H., . . . Lewis, J. W. (2020). Processing of Real-World, Dynamic Natural Stimuli in Autism is Linked to Corticobasal Function. *Autism Res*, *13*(4), 539-549. doi:10.1002/aur.2250
- Werner, S., & Noppeney, U. (2010). Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *J Neurosci*, *30*(7), 2662-2675. doi:10.1523/JNEUROSCI.5091-09.2010
- Willems, R. M., Ozyurek, A., & Hagoort, P. (2007). When language meets action: the neural integration of gesture and speech. *Cereb Cortex*, *17*(10), 2322-2333. doi:10.1093/cercor/bhl141
- Wolf, D., Schock, L., Bhavsar, S., Demenescu, L. R., Sturm, W., & Mathiak, K. (2014). Emotional valence and spatial congruency differentially modulate crossmodal processing: an fMRI study. *Front Hum Neurosci*, *8*, 659. doi:10.3389/fnhum.2014.00659

Zilbovicius, M., Meresse, I., Chabane, N., Brunelle, F., Samson, Y., & Boddaert, N. (2006). Autism, the superior temporal sulcus and social perception. *Trends Neurosci*, 29(7), 359-366.

UNCORRECTED MANUSCRIPT