Brain and Cognition 74 (2010) 97–106

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c

Cortical integration of audio–visual speech and non-speech stimuli

Brent C. Vander Wyk *, Gordon J. Ramsay, Caitlin M. Hudac, Warren Jones, David Lin, Ami Klin, Su Mei Lee, Kevin A. Pelphrey

Yale Child Study Center, Yale University School of Medicine, Yale University, United States

ARTICLE INFO

Article history:
Accepted 12 July 2010
Available online 14 August 2010

Keywords:
Audio–visual
fMRI
Multi-modal processing
Speech

ABSTRACT

Using fMRI we investigated the neural basis of audio–visual processing of speech and non-speech stimuli using physically similar auditory stimuli (speech and sinusoidal tones) and visual stimuli (animated circles and ellipses). Relative to uni-modal stimuli, the different multi-modal stimuli showed increased activation in largely non-overlapping areas. Ellipse-Speech, which most resembles naturalistic audio–visual speech, showed higher activation in the right inferior frontal gyrus, fusiform gyri, left posterior superior temporal sulcus, and lateral occipital cortex. Circle-Tone, an arbitrary audio–visual pairing with no speech association, activated middle temporal gyri and lateral occipital cortex. Circle-Speech showed activation in lateral occipital cortex, and Ellipse-Tone did not show increased activation relative to uni-modal stimuli. Further analysis revealed that middle temporal regions, although identified as multi-modal only in the Circle-Tone condition, were more strongly active to Ellipse-Speech or Circle-Speech, but regions that were identified as multi-modal for Ellipse-Speech were always strongest for Ellipse-Speech. Our results suggest that combinations of auditory and visual stimuli may together be processed by different cortical networks, depending on the extent to which multi-modal speech or non-speech percepts are evoked.

1. Introduction

Although sensory inputs from different modalities are segregated at the periphery, cortical processing must ultimately integrate these inputs. Studies of multi-modal integration have been pursued on different levels of complexity and abstractness. At lower levels, investigators have focused on arbitrary stimuli, such as flashing lights and simple tones. These can be integrated into a unified percept, largely by virtue of their temporal synchronization. At higher levels, the modalities through which a stimulus is presented may still have some temporal relationship, but may also have further associations beyond the simple temporal one. For example, in audio–visual speech the sound of speech and the movement of the lips are related in time, but are also related by phonological knowledge. Multi-modal integration has been shown to support faster response times and better detection rates (Frens, Van Opstal, & Van der Willigen, 1995; Hughes, Reuter-Lorenz, Nozawa, & Fendrich, 1994; Miller, 1982), and improve detection and recognition of speech in noise (Grant & Seitz, 2000; Sumby & Pollack, 1954). There is now a large body of work on the neural basis of multi-modal integration, focussing on one level or the other, but rarely both.

Early investigations of the neural basis of audio–visual integration used electrophysiological recordings of small numbers of cells in the superior colliculus to demonstrate increased activity to very simple co-occurring audio–visual stimuli (King & Palmer, 1985; Meredith & Stein, 1983, 1986), though more recent work has examined cortical cells in the superior temporal sulcus (STS) (Barraclough, Xiao, Baker, Oram, & Perrett, 2005). The development of human neuroimaging methodologies have made it possible to investigate the integration of more complex stimuli in the human cortex. The use of simple audio–visual stimuli has continued to play a role in the understanding of multi-modal integration in the human brain. Calvert, Hansen, Iversen, and Brammer (2001), for example, presented rotating checkerboards paired with bursts of white noise and found integration-related activity in the STS, the superior colliculus, insula, frontal regions, parietal regions, and occipital gyrus. Degerman and colleagues (2007) found greater activity in the inferior and middle frontal gyrus (MFG) and in the temporal cortex when participants attended to the conjunction of multiple modalities as opposed to a single modality. Additionally, integration of moving visual and auditory information activated bilateral superior temporal cortex and the precuneus (Baumann & Greenlee, 2007).

Functional magnetic resonance imaging of human participants has also made it possible to study the neural basis of audio–visual (AV) integration as it affects conscious perception. For instance, Dhamala, Assisi, Jirsa, Steinberg, and Kelso (2007) manipulated

* Corresponding author.
E-mail address: brent.vanderwyk@yale.edu (B.C. Vander Wyk).

0278-2626/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.bandc.2010.07.002

the phase and temporal offset of stimulus tones and lights during an fMRI scan and found that only certain parameters led to a perception of synchrony. This perception was associated with increased activity in the MFG and the superior temporal gyrus (STG). It has been discovered that illusory perceptions can result from multi-modal perception as well. For example, an illusory shift of the perceived location of an auditory source toward a synchronously presented visual stimulus (i.e. a "ventriloquism" effect) is associated with increased activity in the left STS, bilateral parieto-occipital sulcus, and right insula (Bischoff et al., 2007). In a more basic demonstration, the perception, sometimes illusory, of seeing a single flash versus two flashes of light was dependent on whether participants simultaneously heard one or two tones. Activity in V1 to flashes of light was altered by the presence of tones (Watkins, Shams, Josephs, & Rees, 2007; Watkins, Shams, Tanaka, Haynes, & Rees, 2006). The perception of a single flash versus two flashes matched this altered V1 activity, suggesting that sound inputs can affect visual processing on its way to conscious perception at quite an early stage. Crossmodal effects in what was thought to be uni-modal sensory cortex have now been reported a number of times. The peak of the hemodynamic response in primary auditory and visual cortex to multi-modal stimuli as compared with uni-modal stimuli is shifted (Martuzzi et al., 2007). Direct connections between primary auditory and visual cortices have also been implied by functional connectivity analyses of cortical activity during multi-modal stimulus processing. Eckert and colleagues (2008) showed that primary auditory and visual cortices were tightly coupled while at rest and that these regions remained coupled during a visual task while other regions decoupled.

The perception of audio–visual speech, even of meaningless speech such as nonsense syllables or single vowels, entails an additional level of complexity. Audio–visual speech, like the arbitrary stimuli used in the studies described above, exhibits temporal synchrony between the actions of the mouth and its acoustic consequences. Beyond this simple temporal relationship, these speech and mouth movements are associated with one another by way of knowledge about the phonetic structure of language. Meaningful audio–visual speech, such as whole words and phrases, invokes additional associations through lexical, syntactic, or semantic knowledge. Several regions have been shown to have superadditive activation, or simply heightened activation, to meaningful AV speech (Calvert, Campbell, & Brammer, 2000; Calvert et al., 1999) in regions including the right fusiform, lateral occipital cortex, bilateral STS, left MFG, and Heschl's gyrus (HG). Regions in the STS were bilaterally more active to AV speech than uni-modal conditions (Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003). As expected, the activity of these regions is affected by task and stimulus properties. For instance, videos of actors speaking sentences with a variety of emotional expressions elicit greater activation in STS regions compared to audio and video alone (Robins, Hunyadi, & Schultz, 2009).

Few studies have contrasted uni-modal presentation with multi-modal presentation using meaningless speech. However, in a matching paradigm using faces and meaningless speech, there was greater activation to crossmodal matches than to uni-modal matches in the intraparietal sulcus (IPS), superior parietal lobule (SPL), and dorsal premotor area (PMd) (Saito et al., 2005). Superadditive multi-modal effects such as the ones reported above are often found in the context of inverse effectiveness (Stanford & Stein, 2007), wherein a weak stimulus generates a proportionally stronger reaction than does a stronger stimulus. For example, co-presentation of a visual stimulus with an auditory stimulus may produce a weak effect if the auditory stimulus is clearly audible, but a strong effect if the audio signal is degraded. Inverse effectiveness has been demonstrated in speech processing in HG, STS, inferior frontal gyrus (IFG), and medial occipital gyrus (MOG) (Stevenson & James, 2009). Szycik, Tausche, and Munte (2008) also found bilateral STS regions that evinced an interaction between audio–visual synchrony and intelligibility, indexed by amount of noise.

A classic finding in AV speech is the McGurk effect, wherein the perception of auditory speech is altered by the presence of visual information (McGurk & MacDonald, 1976). The co-presentation of incongruent auditory and visual stimuli results in a percept that corresponds to neither auditory nor visual stimulus alone, but rather a fusion that is intermediate between the two. For example, an auditory /ba/ paired with a visual /ga/ is often perceived as a /da/. Neuroimaging studies using McGurk paradigms have found greater activation to incongruent (McGurk-eliciting) stimuli than to congruent stimuli in supramarginal gyrus, inferior parietal lobule and right precentral gyrus (Jones & Callan, 2003), as well as stronger and more posterior left STS activation when the audio component was less intelligible (Sekiyama, Kanno, Miura, & Sugita, 2003). Using whole words, greater activation was found in the left temporal pole and left claustrum to speaking faces paired with McGurk-inducing (synchronous) audio compared to temporally offset audio–video stimuli (Olson, Gatenby, & Gore, 2002). Attention to unsynchronized speech resulted in lower activation in the STS bilaterally (Fairhall & Macaluso, 2009), consistent with a behavioral finding in which a competing attentional demand reduced the McGurk effect (Alsius, Navarra, Campbell, & Soto-Faraco, 2005).

Not all mismatches between audio and visual inputs are integrated into a coherent percept. For example, in a poorly dubbed movie, the auditory speech and the visual mouth movements are typically perceived as completely out of sync. This kind of asynchrony seems to induce increased cortical activation, especially in Broca's area (Ojanen et al., 2005), the left superior medial gyrus (SMG) and right STS (Bernstein, Lu, & Jiang, 2008), although it may be the case that there are distinct regions within the superior temporal cortex that are preferentially activated by either synchronous or asynchronous stimuli (Stevenson, Altieri, Kim, Pisoni, & James, 2010). This specialization within the temporal cortices may be related to the finding that a small amount of temporal misalignment actually promotes AV integration. Single cell recordings in rhesus monkeys have shown visually-modulated responses to auditory information when visual input precedes auditory input by 20–80 ms (Kayser, Petkov, & Logothetis, 2008). For more complex stimuli, such as speech, the optimal offset is longer (Dixon & Spitz, 1980; Grant, van Wassenhove, & Poeppel, 2004; McGrath & Summerfield, 1985). In a direct comparison of temporal synchrony with integration at the perceptual level, Miller and D'Esposito (2005) found that the superior colliculus (SC), anterior insula, and anterior IPS showed more activity to synchronously presented stimuli than offset stimuli. They also found that IFG, HG, and STS were more active when the stimuli were perceived as synchronous, even though they were actually offset.

Overall, there is a great deal of overlap between regions involved in audio–visual integration for speech and non-speech stimuli, especially in the MTG. Reported peaks of activation to speech-related stimuli tend to extend more posteriorly in the temporal lobe in both hemispheres and more anteriorly in the left hemisphere. Perhaps unsurprisingly, there also appears to be a left hemisphere bias in the frontal lobe. However, comparing across studies of different aspects of multi-modal integration is problematic for two reasons. First, the kinds of stimuli used in these experiments to study non-speech and speech integration differ in their complexity. The canonical stimuli for non-speech studies are flashing lights and simple tones, whereas studies of speech-related audio–visual integration involve much more complex speech (meaningless or meaningful) paired with the visual presentation

of speaking lips. Second, no single study has systematically compared these processes.

One aim of the current study is to fill this gap in the literature by using audio stimuli (speech and dynamically varying tones) and visual stimuli (mouth-like ellipses and circles) that are matched on complexity and temporal contingencies. Our experimental design is motivated by key experiments in audio–visual speech perception, which demonstrate that human perceivers are sensitive to audio–visual synchrony between auditory and visual stimuli relating to speech. This is true not only when those stimuli correspond to naturalistic faces and voices, but also when artificial shapes and sounds are co-presented and share drastically reduced information about a common underlying speech source.

It is well known that visual information derived from a talking face can significantly enhance the intelligibility of auditory speech, both in the presence of background acoustic noise (Cotton, 1935; Neely, 1956; Sumby & Pollack, 1954) and when the semantic content of the speech is hard to understand (Arnold & Hill, 2001; Reisberg, McLean, & Goldfield, 1987). Studies of cross-modal integration have focused both on the effect of congruency, when the audio and video stimuli are taken from the same or different utterances (McGurk & MacDonald, 1976; Summerfield & McGrath, 1984), the same or different speakers (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003), and also on the role of temporal synchrony, when the audio and video stimuli are congruent but temporally misaligned (Dixon & Spitz, 1980; McGrath & Summerfield, 1985). All of these studies have found consistent improvements in perception when visual information is used to supplement auditory information, even when the information in the video signal is imperfect or even partially inconsistent with the audio signal (Robert-Ribes, Schwartz, & Escudier, 1995; Summerfield, 1987).

Informed by these findings, other studies have attempted to probe the limits to which auditory and visual information can be degraded or rendered incoherent and still trigger a percept of speech. Experiments using amplitude-modulated tones (Grant, Ardell, Kuhl, & Sparks, 1985; Rosen, Fourcin, & Moore, 1981), band-pass-filtered speech (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995) and sinewave speech analogs (Remez, Rubin, Pisoni, & Carrell, 1981) have demonstrated that phonological information can still be recovered even when much of the complex spectral and temporal structure of speech is destroyed. Similarly, experiments involving Lissajous figures resembling lips (McGrath & Summerfield, 1985) or cartoon analogs of faces (Massarro & Cohen, 1983) have shown that even drastically reduced visual information about a talking face will still enhance the perception of auditory speech.

A consistent finding from all these studies involving distorted or degraded audio–visual speech is that auditory and visual stimuli that are perceived as speech-like when co-presented bimodally may often be perceived as non-speech-like when presented in a single modality (Summerfield, 1979). Thus, in these cases, the speech-like nature of the percept resides crucially in the pairing of common aspects of auditory and visual stimuli taken together, rather than from properties of either stimulus taken in isolation. This immediately suggests the possibility of creating sets of auditory, visual, and audio–visual stimuli that are precisely matched in terms of physical properties, as presented to ear and eye, but which may trigger different cortical responses as speech or non-speech depending on how they are co-presented. In the context of a neuroimaging study, this has the great advantage of allowing us, for the first time, to separate out effects that may arise simply from differences in physical stimulus properties from differences that truly reflect differential activation of cortical networks specific to the integration of speech and non-speech.

Here we exploit this advantage by pairing two sets of video stimuli (circles and ellipses) that share similar visual features, with two sets of audio stimuli (speech and tones) that share similar auditory properties. In doing so, we are able to explore how speech-like percepts may be evoked by combining images that are lip-like or non-lip-like with sounds that resemble or do not resemble speech.

2. Materials and methods

2.1. Participants

Sixteen adults between the ages of 19 and 35 years (mean age = 26 years) were recruited through online ads and flyers. Participant's wakefulness was monitored via a MR compatible camera focused on one eye of the participant. Fourteen participants completed both experimental blocks, and two participants completed one block, but fell asleep during the other. Six participants were female, and all were right-handed. All had normal hearing, and either normal or corrected-to-normal (via MRI compatible lenses) vision.

2.2. Stimuli

Stimulus materials were generated as part of a larger study of audio–visual processing in infants, children, and adults. Two female speakers, both professional actresses, were seated in front of a camera and were recorded acting out scripts involving child-directed caregiver interactions. Four 8-s segments of speech were selected from the set of recordings for use in the present study. The segments consisted of whole sentences in American English and were chosen to exhibit a range of phonetic and rhythmic patterns typical of natural speech. They were also carefully selected so that each token was the same duration. All audio signals were recorded using a high-quality microphone at a sampling rate of 44.1 kHz with 16-bit resolution.

The segments excised directly from the audio signals were used as speech stimuli. To create our non-speech stimuli, the amplitude envelope of each audio segment was first determined by calculating the root-mean-squared (RMS) amplitude of the speech signal averaged over a sliding window, 75 ms in duration, centered at each sample point in the original signal. (Using a 75 ms window length resolves the main syllable peaks and word boundaries in the original utterance, but is guaranteed to eliminate all of the segmental detail at the level of individual glottal cycles.) To control for irrelevant differences in loudness between recordings, the amplitude envelope was normalized by linearly rescaling the values so that the maximum absolute value of each signal was set to unity. Each amplitude envelope was then multiplied with a sine wave of constant frequency 400 Hz, unit amplitude, and identical duration, to obtain a synthetic amplitude-modulated tone that shares many of the temporal properties of the original speech, but destroys all of the spectral detail needed to render it intelligible as speech. The four speech waveforms, amplitude envelopes, and amplitude-modulated tones are shown in Fig. 1.

For each pair of speech and non-speech auditory stimuli, a corresponding pair of video stimuli was also generated, consisting of computer-animated geometric shapes constructed to covary in synchrony with the amplitude envelope of the original audio signal. Using Autodesk Maya 2008, a 320 × 240-pixel bitmap containing a random pattern of gray dots was first created to serve as a neutral background. A single frame was then generated for each sample of the original audio signal by superimposing a three-dimensional image of either a circle or an ellipse on the neutral background image. The overall scaling was chosen so that, on average, the figures generated for all of the utterances filled about two-thirds of the background image. The circles and shapes were generated from radially symmetric tube primitives using a red opaque material lit from the front by a diffuse ambient light source.

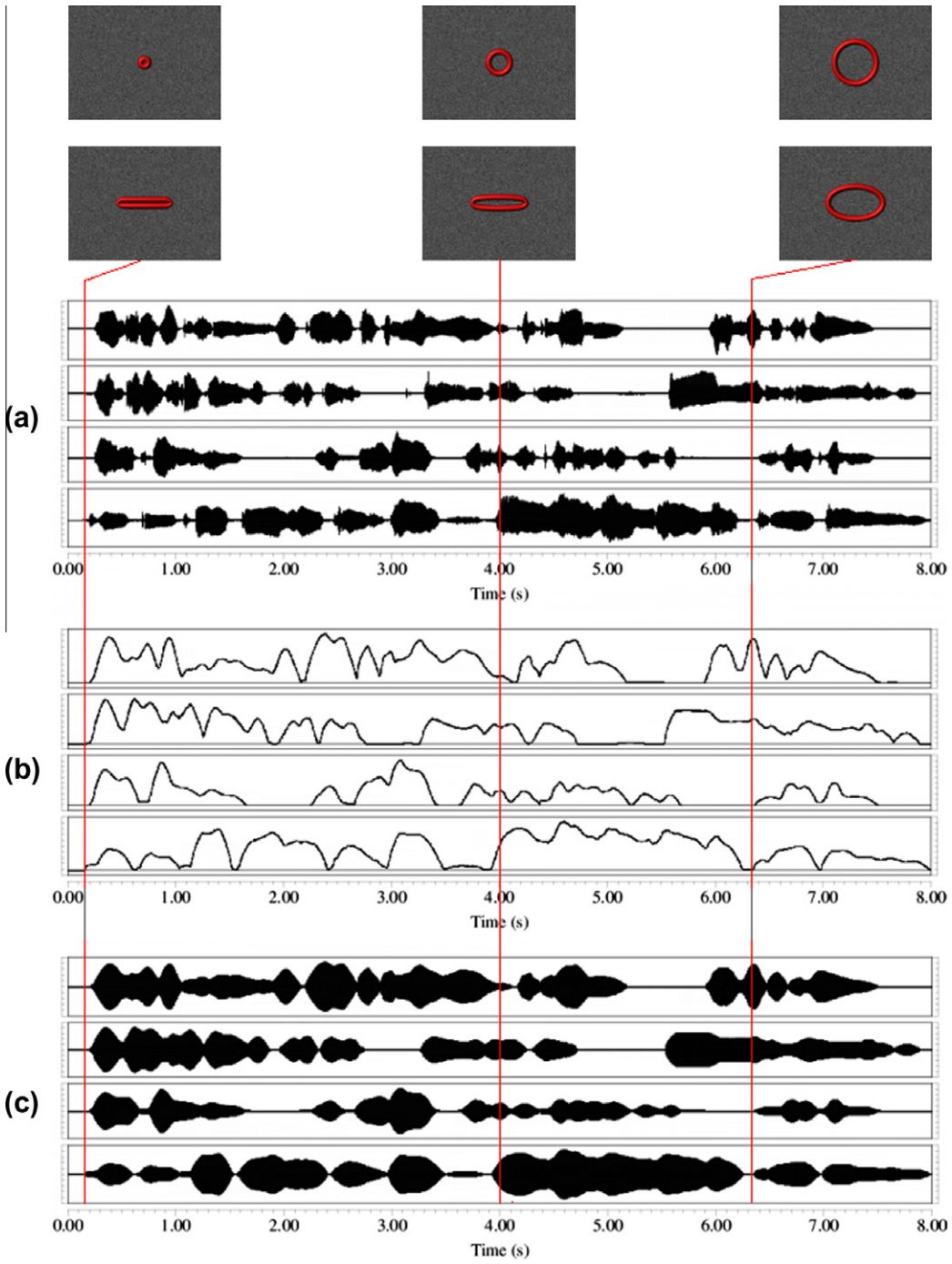

Fig. 1. Stimulus materials, showing (a) original speech waveforms; (b) amplitude envelopes; and (c) amplitude-modulated tones for the four different utterances used to creating the audio–visual stimuli, together with examples of movie stills sampled from the circle and ellipse conditions for the first utterance at three time points.

Examples of the resulting circle and ellipse images taken from the same sample point are shown in Fig. 1. The vertical dimension of the circle and ellipse at each point was chosen so that the total area enclosed by each shape was proportional to the value measured from the amplitude envelope at that point. To make the ellipse appear lip-like, the major axis of the ellipse was scaled to twice the length of the minor axis.

Finally, to create the full set of audio–visual stimuli, each sequence of video frames was down-sampled to a standard video rate of 30 Hz and combined with one of the audio signals using Final Cut Pro 6.0.5, to produce a movie, which was saved to disk in AVI format using the H.264 codec. The two audio stimuli from each recording were combined in this manner with each of the corresponding two video stimuli taken from the same recording. In this way, four audio–visual stimuli were generated and labeled Circle-Tone, Circle-Speech, Ellipse-Tone, and Ellipse-Speech for each of the four recordings, yielding sixteen audio–visual stimuli in total. Visual stimuli were displayed on a rear projection screen viewable by the participant in the scanner via a mirror mounted on the head coil. Audio stimuli were presented over MR compatible headphones (Resonance Technologies, Northridge, CA).

2.3. Study design

The experiment consisted of two 11′40″ functional runs in which participants passively viewed and listened to the stimuli

presented through E-Prime 2.0 software (Psychological Software Tools, Inc., Pittsburgh, PA). Each run was initiated and terminated with 16 s of fixation (white cross on a black background). Stimuli were presented in blocks separated by 12 s of fixation. Each block consisted of two stimuli from the same condition played in succession. The pair of stimuli was always derived from different speech tokens, so the audio and video components did not repeat within a block. At the level of the underlying speech token, presentation order of stimuli was counterbalanced across conditions. Each functional run consisted of three blocks of each condition type. The order of blocks was pseudorandomized across participants.

2.4. Data acquisition

Scanning was performed on a Siemens 3 Tesla Trio scanner (Siemens, Erlangen, Germany) at the Magnetic Resonance Research Center, Yale University School of Medicine. High resolution anatomical images were acquired using a 3D MPRAGE sequence. Whole-brain functional images were acquired using a single-shot, gradient-recalled echo planar pulse sequence (TR = 2000 ms; TE = 25 ms; flip angle = 60°; FOV = 22 cm; image matrix = 64 × 64; voxel size = 3.2 × 3.2 × 3.2 mm; 34 slices) sensitive to blood oxygenation level-dependent (BOLD) contrast.

2.5. Data analysis

Functional data were analyzed with the BrainVoyager QX 10.10.1250 (BrainInnovation, Maastricht, the Netherlands) software package. The first two volumes were discarded to allow for T1 equilibrium. Preprocessing of the functional data included slice scan time correction using cubic spline interpolation, 3D motion correction using trilinear interpolation to correct for small head movements, linear trend removal, and temporal high pass filtering to remove low-frequency non-linear drifts 2 or fewer cycles per time course (2.86×10^{-3} Hz). Functional images were co-registered to each individual's anatomical volume and transformed into Talairach space (Talairach & Tournoux, 1988).

A general linear model (GLM) was used to compute first-level statistics on the z-normalized BOLD signal for each individual. The model time course for each predictor was computed by convolving a gamma function (Boynton, Engel, Glover, & Heeger, 1996) with a boxcar function equal to 1.0 when the condition was present in the experiment and 0.0 otherwise. Using the GLM, parameter estimates for each condition were calculated for each voxel, excluding voxels outside the brain. The results of the first level analysis were entered into second-level random-effects analyses, described below, to account for between-subject variability. In order to account for multiple comparisons, a voxel-level uncorrected threshold was first set to $p < 0.01$. Next, a cluster-size threshold was computed using an iterative Monte Carlo simulation to estimate an acceptable cluster-level false-positive rate (Forman et al., 1995; Goebel, Esposito, & Formisano, 2006). After 5000 iterations, a minimum cluster-size that yielded a false positive rate of $p < 0.01$ was applied to statistical maps.

2.6. Conjunction of contrasts

Regions with multi-modal response properties were identified using a random-effects conjunction analysis. Each multi-modal condition was contrasted against each of the corresponding uni-modal conditions (e.g. Ellipse-Speech versus Ellipse-Only and Ellipse-Speech versus Speech-Only). Only voxels that were significantly active in both contrasts counted as multi-modal. Significance levels were set to $p < 0.01$, corrected for multiple comparisons using a cluster threshold of 8 functional voxels.

2.7. Multi-modal interactions

The conjunction analysis approach is useful for identifying regions that have multi-modal response properties. However, it may obscure similarities in the response of a voxel to different multi-modal stimuli. That is, a voxel may meet the requirements for multi-modality for one stimulus, but not another, even if it has the same response to each. To investigate the response properties to the audio and visual components that made up the multi-modal stimuli, we performed a repeated-measures ANOVA on the beta values from the first order statistical analysis. These values were restricted to the voxels that were found to be significant in the conjunction analysis. One factor corresponded to the visual stimulus type (Circle or Ellipse) and the second factor corresponded to the audio stimulus type (Tone or Speech). In this analysis we were only concerned with multi-modal responses, so the uni-modal activations were not included in the ANOVA.

2.8. Stimulus congruency

Several theories of STS function suggest that it plays a role in amodal matching of stimuli on dimensions other than simple correlations in temporal contingencies in an audio and visual stream (Hocking & Price, 2008). For example, the sound of a duck is congruent with an image of a duck, but not in virtue of shared temporal structure, whereas an image of a dog with the same sound is incongruent. Several studies have found differential brain activity to incongruent and congruent stimuli (Hocking & Price, 2009; Laurienti et al., 2003; Sestieri et al., 2006). A prediction of this hypothesis is that there should be differential activity for the image of a mouth-like stimulus paired with speech (which is congruent) compared to the image of a less mouth-like stimulus paired with speech (which is less congruent). In the current study, Ellipse-Speech and Circle-Speech are well-matched in terms of movement and synchrony, but may differ in their congruency. The former is composed of a congruent mouth-like image and speech, while the latter is composed of a less congruent, less mouth-like image and speech. Therefore, we performed a direct contrast of Ellipse-Speech and Circle-Speech across the whole brain.

3. Results

3.1. Conjunction of contrasts

The result of this analysis was 14 clusters of voxels distributed across the brain. There was one cluster for Circle-Speech, six for Circle-Tone, and seven for Ellipse-Speech. No clusters were found for Ellipse-Tone. The Talairach coordinates of the peak voxel for these regions, as well the extent and anatomical location, is reported in Table 1. The regions are also plotted in Fig. 2.

3.2. Multi-modal interactions within conjunction regions

The multi-modal regions demonstrating significant interaction effects are listed in Table 2 and also plotted on a display brain in Fig. 3, along with plots of average beta values in order to interpret the interaction effect. It is apparent that the interaction manifests in different patterns among the brain regions. We used paired, two-tailed t-tests with a Bonferroni correction to evaluate specific comparisons. The region in the IFG showed significant differentiation between tone and speech, $t(15) = 2.70$, $p < 0.05$ when paired with an ellipse, but not when paired with a circle. The regions in

Table 1
Peak voxels for conjunction analysis.

Condition	Location	Hemisphere	Talairach coordinates			Extent	Peak *t*-score
			X	*Y*	*Z*		
Circle-Speech	Middle occipital gyrus	R	25	−84	−4	20	4.82
Circle-Tone	Middle Temporal gyrus	R	53	−34	8	14	4.01
	Superior temporal gyrus	R	46	−42	8	15	4.00
	Fusiform gyrus	R	38	−39	−16	8	4.00
	Lingual gyrus	R	25	−78	−4	33	3.82
	Lingual gyrus	L	−23	−78	−4	26	5.04
	Middle temporal gyrus	L	−58	−35	11	17	6.30
Ellipse-Speech	Middle frontal gyrus	R	46	9	32	13	3.82
	Fusiform gyrus	R	34	−59	−16	41	5.48
	Middle occipital gyrus	R	20	−87	−6	36	6.53
	Medial frontal gyrus	L	−6	45	−8	8	4.75
	Lingual gyrus	L	−14	−87	−8	45	5.08
	Fusiform gyrus	L	−37	−63	−15	21	5.91
	Middle temporal gyrus	L	−47	−60	7	10	4.05

Clusters found to be significant to a conjunction of contrasts of the multi-modal condition to respective uni-modal conditions ($p < 0.01$, corrected for multiple comparisons using cluster thresholds). Extent of regions in specified in functional voxels (3 mm^3). Anatomical labels were derived from the Talairach.org database (Lancaster et al., 2000).

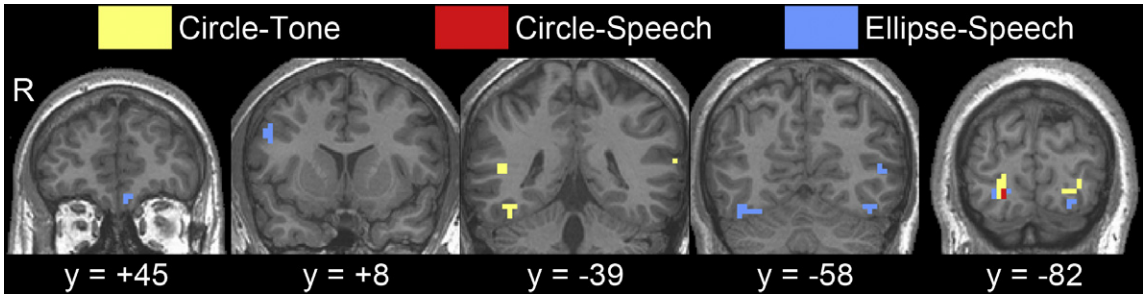

Fig. 2. Regions showing significance ($p < 0.01$, corrected for multiple comparisons) in a conjunction of contrasts of the multi-modal condition to single modal conditions. Yellow = Circle-Tone > Circle-Only and Circle-Tone > Tone-Only; Red = Circle-Speech > Circle-Only and Circle-Speech > Speech-Only; Blue = Ellipse-Speech > Ellipse-Only and Ellipse-Speech > Speech-Only.

Table 2
Peak voxels for interaction effects among conjunction regions.

Location	Hemisphere	Talairach coordinates			Extent
		X	*Y*	*Z*	
Inferior frontal gyrus	R	46	9	29	9
Superior temporal gyrus	R	43	−32	6	12
Fusiform gyrus	R	34	−66	−17	21
Lingual gyrus	L	−14	−87	−1	23
Fusiform gyrus	L	−34	−63	−15	5
Superior temporal gyrus	L	−61	−39	12	5

Clusters found to have a significant interaction effect among audio and visual factors ($p < 0.01$, corrected for multiple comparisons using cluster thresholds). Extent of regions in specified in functional voxels (3 mm^3). Anatomical labels were derived from the Talairach.org database (Lancaster et al., 2000).

the right and left STG showed a significant difference between circle and ellipse when paired with a tone, $t(15) = 4.54$, $p < 0.001$, and $t(15) = 4.65$, $p < 0.001$, respectively, but showed no difference between circle and ellipse when paired with speech. The right and left fusiform regions both showed significantly greater activity to speech when paired with an ellipse than when paired with a circle, $t(15) = 5.55$, $p < 0.001$ and $t(15) = 4.59$, $p < 0.001$, respectively. However, there was no significant difference in activation to tones, although the right fusiform did show a trend for Circle-Tone greater than Circle-Speech, $t(15) = 2.22$, $p = 0.08$. The left MOG exhibited a very complex pattern. Activation to Ellipse-Speech was greater than Ellipse-Tone, $t(15) = 4.20$, $p < 0.01$, while activation to Circle-Tone was greater than Circle-Speech, $t(15) = 3.06$, $p < 0.05$.

3.3. Stimulus matching

Shown in Fig. 4, one region in the right FFG (peak voxel at Talairach coordinates 35, −60, −9) showed significantly greater activation to Ellipse-Speech than Circle-Speech, $p < 0.0001$, corrected for multiple comparisons using a cluster threshold ($k = 8$).

4. Discussion

In the current study, we used stimuli with matched temporal synchrony to compare uni-modal presentations of auditory stimuli and visual stimuli to the multi-modal stimuli formed by their conjunction. We controlled for key aspects of both the auditory complexity of the speech and non-speech sounds and the visual complexity of the mouth-like (ellipse) and non-mouth-like (circle) images. This allowed us to explore the cortical processing of both arbitrarily paired multi-modal stimuli (Circle-Tone), non-arbitrarily associated multi-modal stimuli (Ellipse-Speech) which resembled the stimulus in visual speech perception, and combinations of intermediate stimuli (Circle-Speech and Ellipse-Tone).

In the comparison of uni-modal to multi-modal stimuli, we found several regions that were more active to specific multi-modal stimuli than to the constituent uni-modal stimuli, using a conjunction definition. Consistent with previous audio–visual studies reporting activity in lateral occipital cortex (Bischoff et al., 2007; Calvert et al., 1999, 2000, 2001; Stevenson & James, 2009), we found multi-modal responses in bilateral middle occipital and lingual gyri. Apart from Ellipse-Tone, overlapping clusters of activation for every multi-modal condition were found in this region.

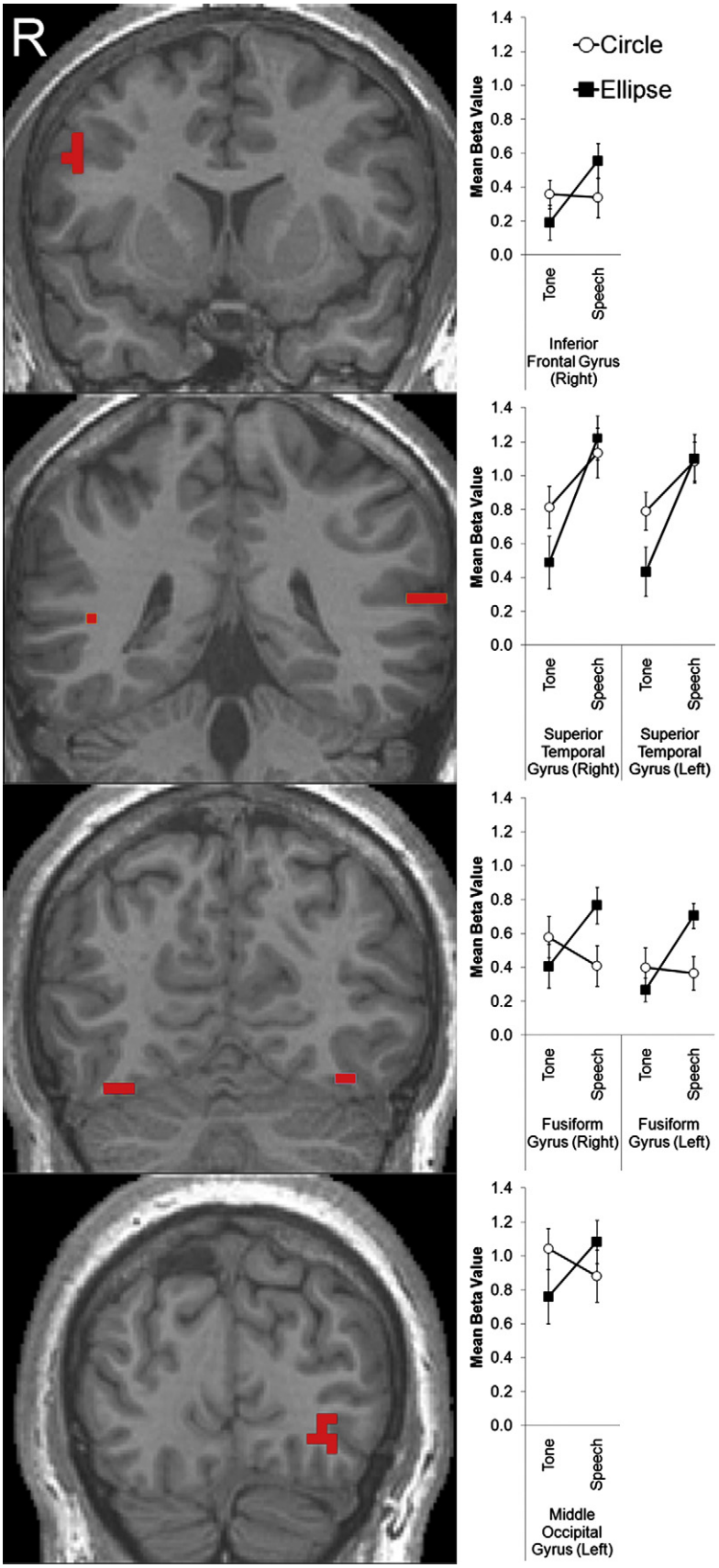

Fig. 3. Regions showing a significant interaction ($p < 0.01$, corrected for multiple comparisons) among multi-modal conditions and mean beta plots.

Circle-Tone stimuli were associated with bilateral activation in MTG and the right anterior FFG. Replicating previous studies of audio–visual speech perception, the Ellipse-Speech condition was associated with activation in the left posterior MTG. Additionally,

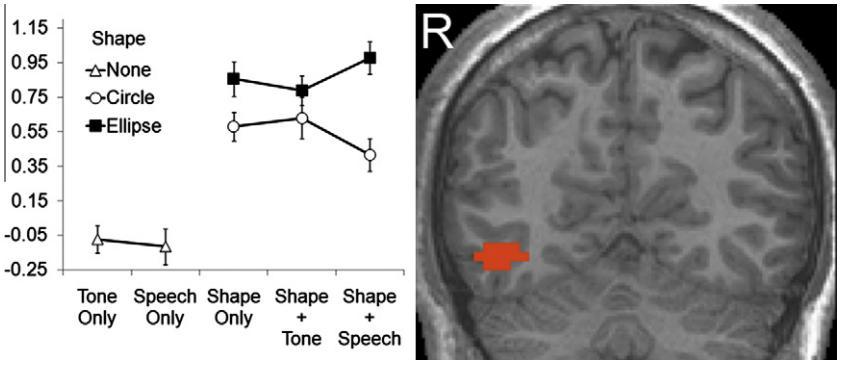

Fig. 4. Right ventral-temporal region showing greater activity ($p < 0.001$, corrected for multiple comparisons) to Ellipse-Speech than Circle-Speech and a plot of mean betas by condition.

we found a cluster in the right MFG and clusters bilaterally in FFG. By itself, the fact that regions responding to the specific multi-modal stimuli were largely non-overlapping, suggests that neural regions supporting integration of arbitrary AV stimuli based on lower-level temporal dynamics are distinct from those supporting the integration of speech and mouth stimuli. Given that (1) Ellipse-Speech does entail temporal correspondences between visual and audio information and (2) several papers have reported responses in MTG to AV speech (Fairhall & Macaluso, 2009; Kang et al., 2006; Szycik et al., 2008), this result was unexpected. The latter point could be partially explained by the fact that most studies that found activation in the MTG compared congruent and incongruent speech, as opposed to comparing multi-modal to uni-modal stimuli.

However, a secondary analysis which directly compared the responses of multi-modal regions to each of the multi-modal conditions revealed that a subset of voxels within bilateral MTG, previously identified as multi-modal to Circle-Tone, actually showed a larger response to the multi-modal speech conditions. Thus, these areas are quite strongly responsive to the multi-modal speech conditions, and in all likelihood, were not found in the conjunction analysis for Ellipse-Speech simply because the speech response was so high. Several regions found in the conjunction analysis to be active to Ellipse-Speech also exhibited significant interactions in the secondary analysis, including the IFG and the FFG. The IFG did not differentiate between the Circle-Tone and Circle-Speech conditions, but did differentiate between Ellipse-Speech and Ellipse-Tone. The fusiform regions also did not differentiate between Circle-Tone and Circle-Speech, but had a larger response to Ellipse-Speech than Circle-Speech, which may indicate a modulation of processing based upon the similarity to naturalistic audio–visual speech. Unlike the multi-modal regions in the MTG found to be more active to Speech conditions, these exhibited patterns of activation consistent with integration of specific audio and visual stimuli, namely mouth-like images and speech sounds.

A direct comparison of Ellipse-Speech to Circle-Speech revealed one large region in the right FFG (Fig. 4). Among the multi-modal stimulus conditions, activation was highest during the Ellipse-Speech condition, while activation during Circle-Speech was lowest relative to the other conditions that included visual input. Portions of the fusiform have long been associated with face processing, and activation in the right fusiform specifically has been associated with increased speech reading ability (Capek et al., 2008). While the fusiform has not typically been identified as multi-modal, this study as well as several other recent reports (Hertrich, Dietrich, & Ackermann, 2010; Stevenson, Kim, & James, 2009) suggest that it may, indeed, be importantly related to audio–visual speech, although it should be noted that this region was also highly active to uni-modal visual conditions.

In this study we focused on multi-modal processing of temporally synchronized audio–visual speech and non-speech stimuli. However, several other kinds of multi-sensory processing have been identified. One is the integration of other sensory modalities, for example, haptic input. Interestingly, a recent study looking at audio, visual, and haptic modalities found distinct networks for integrating each pair (Stevenson et al., 2009). Similarly, another line of research investigating the integration of audio–visual information across space, as opposed to through time, found distinct brain regions for each kind of integration (Macaluso, George, Dolan, Spence, & Driver, 2004). Thus, the networks we have identified here may be specific to the temporal integration of audio–visual information.

Finally, researchers have also identified the multimodal semantic congruency (e.g. a picture of a dog paired with the sound of a dog barking) as an important aspect of multi-modal processing. A large body of work on the neural basis of this kind of integration has indentified a network of regions that partially overlap with those identified in the current study, including the STS, the STG, and lateral occipital cortices (Alpert, Hein, Tsai, Naumer, & Knight, 2008; Beauchamp, Lee, Argall, & Martin, 2004; Hein et al., 2007; Hocking & Price, 2009; Sestieri et al., 2006; Stevenson, Geoghegan, & James, 2007; Taylor, Moss, Stamatakis, & Tyler, 2006; Werner & Noppeney, 2009). Because of this overlap, we cannot make the argument that the regions we identified are specific to the multi-modal integration of speech, per se. A direct comparison between synchrony and congruency might identify where these networks overlap and where they diverge, however, methodological differences between how experiments typically measure and control temporal synchronicity and semantic congruency make this a challenge. Nevertheless, future research is required to understand the relationship between these kinds of multi-modal processing in the brain.

5. Conclusion

Using sets of auditory, visual, and audio–visual stimuli that were matched in terms of their physical properties, this experiment demonstrated differential activation of cortical networks involved in the integration of speech and non-speech stimuli. In particular, the bilateral STG and right MOG exhibited greater activation to all speech conditions than to tone conditions, while the bilateral FFG and right MTG showed greater activation to the conjunction Ellipse-Speech, specifically. These results indicate that these regions are integral components of a speech processing network and suggest a common system of non-overlapping regions involved in multi-modal speech perception. Accordingly, these brain regions may have a role in modulating increasingly complex poly-sensory stimulation.

6. Funding

This research was supported by the John Merck Scholars Fund, the Simons Foundation, and the National Institute of Mental Health (R01-MH083712).

Acknowledgments

We would like to thank our participants for their valuable time and effort, as well as our colleagues Todd Constable, Emily Lechner, and the Magnetic Resonance Research Center staff for their valuable assistance with this research.

References

Alpert, G. F., Hein, G., Tsai, N., Naumer, M. J., & Knight, R. T. (2008). Temporal characteristics of audiovisual information processing. *Journal of Neuroscience, 28*, 5344–5349.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology, 15*, 839–843.

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speech reading advantage when speech is clearly audible and intact. *British Journal of Psychology, 92*, 339–355.

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W., & Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience, 17*, 377–391.

Baumann, O., & Greenlee, M. W. (2007). Neural correlates of coherent audiovisual motion perception. *Cerebral Cortex, 17*, 1433–1443.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41*, 809–823.

Bernstein, L. E., Lu, Z. L., & Jiang, J. T. (2008). Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Research, 1242*, 172–184.

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*, 578–586.

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience, 16*, 4207–4221.

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*, 2619–2623.

Calvert, G. A., Campbell, R., & Brammer, 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*, 427–438.

Capek, C. M., MacSweeney, M., Woll, B., Waters, D., McGuire, P. K., David, A. S., et al. (2008). Cortical circuits for silent speechreading in deaf and hearing people. *Neuropsychologia, 46*, 1233–1241.

Cotton, J. C. (1935). Normal 'visual hearing'. *Science, 82*, 592–593.

Degerman, A., Rinne, T., Pekkola, J., Autti, T., Jääskeläinen, I. P., Sams, M., et al. (2007). Human brain activity associated with audiovisual perception and attention. *Neuroimage, 34*, 1683–1691.

Dhamala, M., Assisi, C. G., Jirsa, V. K., Steinberg, F. L., & Kelso, J. A. S. (2007). Multisensory integration for timing engages different brain networks. *Neuroimage, 34*, 764–773.

Dixon, N. F., & Spitz, L. T. (1980). The detection of audiovisual desynchrony. *Perception, 9*, 719–721.

Eckert, M. A., Kamdar, N. V., Chang, C. E., Beckmann, C. F., Greicius, M. D., & Menon, V. (2008). A cross-modal system linking primary auditory and visual cortices: Evidence from intrinsic fMRI connectivity analysis. *Human Brain Mapping, 29*, 848–857.

Fairhall, S. L., & Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *European Journal of Neuroscience, 29*, 1247–1257.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magnetic Resonance Medicine, 33*, 636–647.

Frens, M. A., Van Opstal, A. J., & Van der Willigen, R. F. (1995). Spatial and temporal factors determine auditory–visual interactions in human saccadic eye movements. *Perception & Psychophysics, 57*, 802–816.

Goebel, R., Esposito, F., & Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Human Brain Mapping, 27*, 392–401.

Grant, K. W., Ardell, L. H., Kuhl, P. K., & Sparks, D. W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *Journal of the Acoustic Society of America, 77*, 671–676.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustic Society of America, 108*, 1197–1208.

Grant, K. W., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. *Speech Communication, 44*, 43–53.

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. *Journal of Neuroscience, 27*, 7881–7887.

Hertrich, I., Dietrich, S., & Ackermann, H. (2010). Cross-modal interactions during perception of audiovisual speech and non-speech signals: An fMRI study. *Journal of Cognitive Neuroscience*. doi:10.1162/jocn.2010.21421.

Hocking, J., & Price, C. J. (2008). The role of the posterior superior temporal sulcus in audiovisual processing. *Cerebral Cortex, 18*, 2439–2449.

Hocking, J., & Price, C. J. (2009). Dissociating verbal and nonverbal audiovisual object processing. *Brain Language, 108*(2), 89–96.

Hughes, H. C., Reuter-Lorenz, P. A., Nozawa, G., & Fendrich, R. (1994). Visual–auditory interactions in sensorimotor processing: Saccades versus manual responses. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 131–153.

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *Neuroreport, 14*, 1129–1133.

Kamachi, M., Hill, H., Lander, L., & Vatikiotis-Bateson, E. (2003). 'Putting the face to the voice': Matching identity across modality. *Current Biology, 13*, 1709–1714.

Kang, E., Lee, D. S., Kang, H. J., Hwang, C. H., Oh, S. H., Kim, C. S., et al. (2006). The neural correlates of cross-modal interaction in speech perception during a semantic decision task on sentences: A PET study. *Neuroimage, 32*, 423–431.

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex, 18*, 1560–1574.

King, A. J., & Palmer, A. R. (1985). Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Experimental Brain Research, 60*, 492–500.

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., Kochunov, P. V., Nickerson, D., Mikiten, S. A., & Fox, P. T. (2000). Automated Talairach Atlas labels for functional brain mapping. *Human Brain Mapping, 10*, 120–131.

Laurienti, P. J., Wallace, M. T., Maldjian, J. A., Susi, C. M., Stein, B. E., & Burdette, J. H. (2003). Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices. *Human Brain Mapping, 19*, 213–223.

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: A PET study. *NeuroImage, 21*(2), 725–732.

Martuzzi, R., Murray, M. M., Michel, C. M., Thiran, J. P., Maeder, P. P., Clarke, S., et al. (2007). Multisensory interactions within human primary cortices revealed by BOLD dynamics. *Cerebral Cortex, 17*, 1672–1679.

Massarro, D. W., & Cohen, M. M. (1983). Speech perception in perceivers with hearing loss: Synergy of multiple modalities. *Journal of Speech and Language Hearing Research, 42*, 21–41.

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio–visual speech recognition by normal-hearing adults. *Journal of the Acoustic Society of America, 77*, 678–685.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 265*, 746–748.

Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science, 221*, 389–391.

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology, 56*, 640–662.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14*, 247–279.

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience, 25*, 5884–5893.

Neely, K. K. (1956). Effect of visual factors on the intelligibility of speech. *Journal of the Acoustic Society of America, 28*, 1275–1277.

Ojanen, V., Mottonen, R., Pekkola, J., Jaaskelainen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage, 25*, 333–338.

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. *Cognitive Brain Research, 14*, 129–138.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A speechreading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). London: Lawrence Erlbaum Associates.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*, 947–949.

Robert-Ribes, J., Schwartz, J.-L., & Escudier, P. (1995). A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review, 9*, 81–104.

Robins, D. L., Hunyadi, E., & Schultz, R. T. (2009). Superior temporal activation in response to dynamic audio–visual emotional cues. *Brain and Cognition, 69*, 269–278.

Rosen, S. M., Fourcin, A. J., & Moore, B. C. J. (1981). Voice pitch as an aid to lipreading. *Nature, 291*, 150–152.

Saito, D. N., Yoshimura, K., Kochiyama, T., Okada, T., Honda, M., & Sadato, N. (2005). Cross-modal binding and activated attentional networks during audio–visual speech integration: A functional MRI study. *Cerebral Cortex, 15*, 1750–1760.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory–visual speech perception examined by fMRI and PET. *Neuroscience Research, 47*, 277–287.

Sestieri, C., Di Matteo, R., Ferretti, A., Del Gratta, C., Caulo, M., Tartaro, A., et al. (2006). 'What' versus 'Where' in the audiovisual domain: An fMRI study. *Neuroimage, 33*, 672–680.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303–304.

Stanford, T. R., & Stein, B. E. (2007). Superadditivity in multisensory integration: Putting the computation in context. *Neuroreport, 18*, 787–792.

Stevenson, R. A., Altieri, N. A., Kim, S., Pisoni, D. B., & James, T. W. (2010). Neural processing of asynchronous audiovisual speech perception. *NeuroImage, 49*, 3308–3318.

Stevenson, R. A., Geoghegan, M. L., & James, T. W. (2007). Superadditive BOLD activation in superior temporal sulcus with threshold non-speech objects. *Experimental Brain Research, 179*, 85–95.

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*, 1210–1223.

Stevenson, R. A., Kim, S., & James, T. W. (2009). An additive-factors design to disambiguate neuronal and areal convergence: Measuring multisensory interactions between audio, visual, and haptic sensory streams using fMRI. *Experimental Brain Research, 198*(2), 183–194.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustic Society of America, 26*, 212–215.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica, 36*, 314–331.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio–visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–51). London: Lawrence Erlbaum Associates.

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio–visual incompatibility in the perception of vowels. *Journal of Experimental Psychology, 36*(A), 51–74.

Szycik, G. R., Tausche, P., & Munte, T. F. (2008). A novel approach to study audiovisual integration in speech perception: Localizer fMRI and sparse sampling. *Brain Research, 1220*, 142–149.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain: 3-Dimensional proportional system: An approach to cerebral imaging*. Thieme: Stuttgart.

Taylor, K. I., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2006). Binding crossmodal object features in perirhinal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 8239–8244.

Watkins, S., Shams, L., Josephs, O., & Rees, G. (2007). Activity in human V1 follows multisensory perception. *Neuroimage, 37*, 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*, 1247–1256.

Werner, S., & Noppeney, U. (2009). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral Cortex*. doi:10.1093/cercor/bhp248 Retrieved 29.06.10 [Advance online publication].

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex, 13*, 1034–1043.