Published in final edited form as:

Brain Cogn. 2010 November ; 74(2): . doi:10.1016/j.bandc.2010.07.002.

Cortical Integration of Audio-Visual Information

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

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

Abstract

We investigated the neural basis of audio-visual processing in speech and non-speech stimuli. Physically identical auditory stimuli (speech and sinusoidal tones) and visual stimuli (animated circles and ellipses) were used in this fMRI experiment. Relative to unimodal stimuli, each of the multimodal conjunctions showed increased activation in largely non-overlapping areas. The conjunction of Ellipse and Speech, which most resembles naturalistic audiovisual speech, showed higher activation in the right inferior frontal gyrus, fusiform gyri, left posterior superior temporal sulcus, and lateral occipital cortex. The conjunction of Circle and Tone, an arbitrary audio-visual pairing with no speech association, activated middle temporal gyri and lateral occipital cortex. The conjunction of Circle and Speech showed activation in lateral occipital cortex, and the conjunction of Ellipse and Tone did not show increased activation relative to unimodal stimuli. Further analysis revealed that middle temporal regions, although identified as multimodal only in the Circle-Tone condition, were more strongly active to Ellipse-Speech or Circle-Speech, but regions that were identified as multimodal 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 speech or non-speech percepts are evoked.

Keywords

audio-visual; fMRI; multi-modal processing

Although sensory inputs from different modalities are segregated at the periphery, ultimately cortical processing reintegrates these inputs. Studies of multimodal integration have been pursued on different strata. 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, and support faster response times and better detection rates than unsynchronized controls (Miller 1982; Hughes et al. 1994; Frens et al. 1995). 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. Integration at this level also has behavioral implications: improving detection and recognition of speech in noise (Sumby and Pollack 1954; Grant and Seitz 2000) or altering its perception (McGurk and MacDonald 1976). The cortical basis of multimodal integration within each of these levels has been studied extensively. However, studies that make up the now large body of

*Please address correspondence to: Kevin A. Pelphrey, Ph.D., Yale Child Study Center, Yale University School of Medicine, Yale University, 230 South Frontage Road, New Haven, CT 06520, Telephone: (203) 393-6871, Fax: (203) 737-4197, kevin.pelphrey@yale.edu.

work illuminating the neural basis of multimodal integration tend to focus 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 coincident audio-visual stimuli (Meredith and Stein 1983; King and Palmer 1985; Meredith and Stein 1986), though more recent work has examined cortical cells in the superior temporal sulcus (STS) (Barracough et al. 2005). The development of human neuroimaging methodologies 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 understanding multimodal integration in the human brain. Calvert et al. (2001) 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. This study also established the use of superadditivity (meaning that the response for the multimodal condition is greater than the sum of the responses for each unimodal condition) as a standard for evaluating whether activity to multimodal stimuli is related to integration, although this standard has been challenged recently (Laurienti et al. 2005; Stanford and Stein 2007). Greater activity was found in the inferior and middle frontal gyrus (MFG) and the temporal cortex when participants attended to the conjunction of multiple modalities as opposed to a single modality (Degerman et al. 2007). Integration of moving visual and auditory information activated bilateral superior temporal cortex and the precuneus (Baumann and 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 and colleagues (2007) manipulated the phase and offset of tones and lights 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). Illusory perceptions can also result from multimodal perception. 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. Moreover, activity in V1 to flashes of light was altered by the presence of tones (Watkins et al. 2006; Watkins et al 2007). The perception of a single flash versus two flashes matched the 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 unimodal sensory cortex have now been reported a number of times. The peak of the hemodynamic response in primary auditory and visual cortex to multimodal stimuli as compared with unimodal 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 multimodal stimulus processing. Primary auditory and visual cortices were tightly coupled while at rest, and these regions remained coupled during a visual task while other regions decoupled (Eckert et al. 2008).

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 et al. 1999; Calvert et al. 2000) in regions including the right fusiform, lateral occipital cortex, bilateral STS, left MFG, and Heschl's gyrus (HG). Regions in STS bilaterally were more active to AV speech than unimodal conditions (Wright et al. 2003). As expected, the activity of these regions is affected by task and stimulus properties. Using videos of actors speaking sentences with a variety of emotional expressions, greater activation was found in STS regions compared to audio and video alone (Robins et al. 2009). Few studies have contrasted unimodal presentation with multimodal presentation using meaningless speech. However, in a matching paradigm using faces and meaningless speech, there was greater activation to crossmodal matches than to unimodal matches in the intraparietal sulcus (IPS), superior parietal lobule (SPL), and dorsal premotor area (PMd) (Saito et al. 2005). Superadditive multimodal effects such as the ones reported above are often found in the context of inverse effectiveness (Stanford and 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 and James 2009). Another study found bilateral STS regions that evinced an interaction between audio-visual synchrony and intelligibility, indexed by amount of noise (Szyck et al. 2008).

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 and MacDonald 1976), so that 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 and Callan 2003), as well as stronger and more posterior left STS activation when the audio component was less intelligible (Sekiya et al. 2003). Using whole words, there was greater activation 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 et al. 2002). Attention to unsynchronized speech resulted in lower activation in the STS bilaterally (Fairhall and Macaluso 2009) consistent with a behavioral finding in which a competing attentional demand reduced the McGurk effect (Alsus et al. 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 dyssynchrony seems to induce increased cortical activation, especially in Broca's area (Ojanen et al. 2005) and the left superior medial gyrus (SMG) and right STS (Bernstein et al. 2007). However, it should be noted that a small amount of temporal misalignment is useful and possibly necessary in AV integration. Single cell recording in rhesus monkeys demonstrated visually-modulated responses to auditory information when auditory input is preceded by visual input by 20–80 ms (Kayser et al. 2008), and for more complex stimuli, such as speech, the optimal offset is longer (Dixon and Spitz 1980; McGrath and Summerfield, 1985; Grant et al. 2004). 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, but that IFG, HG, and STS were more active when the stimuli were perceived as synchronous, 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. There also appears to be a left hemisphere bias in the frontal lobe. However, comparing across studies of different aspects of multimodal 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 study has systematically compared these processes in a single study.

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 audiovisual speech perception, demonstrating that human perceivers are sensitive to audiovisual synchrony between auditory and visual stimuli relating to speech, not only when those stimuli correspond to naturalistic faces and voices, but also when artificial shapes and sounds are co-presented that 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; Sumby and Pollack 1954; Neely 1956) or when the semantic content of the speech is hard to understand (Reisberg et al. 1987; Arnold and Hill 2001). 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 and MacDonald 1976; Summerfield and McGrath 1984), the same or different speakers (Kamachi et al. 2003), and also on the role of temporal synchrony, when the audio and video stimuli are congruent but temporally misaligned (Dixon and Spitz 1980; McGrath and 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 (Summerfield 1987; Robert-Ribes et al. 1995).

Informed by these findings, other studies have attempted to probe the limits to which auditory and visual information can be degraded or rendered incoherent while still triggering a percept of speech. Experiments using amplitude-modulated tones (Rosen et al. 1981; Grant et al. 1985), bandpass-filtered speech (Shannon et al. 1995) and sinewave speech analogs (Remez et al. 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 and Summerfield 1985) or cartoon analogs of faces (Massaro and 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 audiovisual speech is that audio and video 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 audiovisual stimuli that are precisely

matched in terms of the physical properties of the stimulus materials, 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 potential effects of differences in brain activation that may arise simply from differences in physical stimulus properties across experimental conditions 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 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.

Materials and Methods

Participants

Nineteen adults between the ages of 19 and 35 years (mean age = 26 years) were recruited through online ads and flyers. Usable data was collected from 14 participants who completed both experimental blocks, and 2 participants who completed only a single block. Six participants were female. All were right-handed. All had normal hearing, and either normal or corrected-to-normal (via MRI compatible lenses) vision.

Stimuli

Stimulus materials were generated as part of a larger study of audiovisual processing in infants, children, and adults. Two female speakers, both professional actresses, were seated in front of a camera and recorded acting out scripts involving child-directed caregiver interactions. Four eight-second 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, chosen to exhibit a range of phonetic and rhythmic patterns typical of natural speech, and were also carefully selected so that the beginning and end of the speech in each segment coincided in time, with any naturally-occurring pauses distributed randomly across segments. 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 provided our 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 Figure 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, choosing the vertical dimension of the shape so that the total area enclosed by the 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. The overall scaling was chosen so that the figures generated for all of the utterances on average filled about two-thirds of the background image. To make the animations visually attractive, the circles and shapes were generated from radially symmetric tube primitives rather than simple line drawings, using a red opaque material lit from the front by a diffuse ambient light source. Examples of the resulting circle and ellipse images taken from the same sample point are shown in Figure 1.

Finally, to create the full set of audiovisual 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. Each of the two audio stimuli from each recording was combined in this manner with each of the corresponding two video stimuli taken from the same recording to generate four audiovisual stimuli, labeled Circle-Tone, Circle-Speech, Ellipse-Tone, and Ellipse-Speech, for each of the four recordings, yielding sixteen audiovisual stimuli in total. Visual stimuli were displayed on a rear projection screen viewable by the participant via a mirror mounted on the head coil. Audio stimuli were presented over MR compatible headphones. Audibility of stimuli was checked for each participant prior to the experiment.

Study Design

The experiment consisted of two 11:40 functional runs presented through E-Prime 2.0 software (Psychological Software Tools, Inc., Pittsburgh, PA). Each run was initiated and terminated with 16 seconds of fixation (white cross on a black background). Stimuli were presented in blocks separated by 12 seconds 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 3 blocks of each condition type. The order of blocks was pseudorandomized across participants.

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. T1-weighted anatomical images were acquired using an MPRAGE sequence (TR = 2530 ms; TE = 3.34 ms; FOV = 25.6 cm; image matrix = 64²; 1 × 1 × 1 mm). 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 = 64²; voxel size = 3.2 × 3.2 × 3.2 mm; 34 slices) sensitive to blood oxygenation level-dependent (BOLD) contrast.

Data Analysis

Functional data were analyzed with the BrainVoyager QX 10.10.1250 (BrainInnovation, Maastricht, the Netherlands) software package. The first 2 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 and 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 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 inter-subject variability and to allow generalizations to the population level. In order to account for multiple comparisons, first a voxel-level uncorrected threshold was set to $p < 0.01$. Then 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 et al. 2006). After 5000 iterations, a minimum cluster-size ($k = 8$ voxels) that yielded a false positive rate of 1% was applied to statistical maps.

Conjunction of Contrasts

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

Multimodal Interactions

The conjunction analysis approach is useful for identifying regions that have some measure of multimodal response properties. However, it may obscure similarities in the response of a voxel to different multimodal stimuli. That is, a voxel may meet the requirements for multimodality 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 multimodal stimuli, we performed a repeated-measures ANOVA on the beta values from the first order statistical analysis, restricted to those 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 multimodal responses, so the unimodal parameters were not included in the ANOVA.

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 and Price 2008). For example, the sound of a duck is congruent with an image of a duck, but not by 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 (Laurienti et al. 2003; Sestieri et al. 2006; Hocking and Price 2009). 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 incongruent). In the current study, Ellipse-Speech and Circle-Speech are well-matched in terms of movement and synchrony, but differ in their congruency. The former is composed

of a congruent mouth-like image and speech, while the former is composed of an incongruent, less mouth-like image and speech. Therefore, we performed a direct contrast of Ellipse-Speech and Circle-Speech across the whole brain.

Results

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 region, as well the extent and anatomical location, is reported in Table 1. The regions are also plotted in Figure 2.

Multimodal Interactions within Conjunction Regions

The multimodal regions demonstrating significant interaction effects are listed in Table 2 and also plotted on a display brain in Figure 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 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, $t(15) = 3.06$, $p < 0.05$.

Stimulus Matching

Shown in Figure 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 the cluster thresholding technique described previously.

Discussion

In the current study, we used stimuli with matched temporal synchrony to compare unimodal presentations of auditory stimuli and visual stimuli to the multimodal 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 multimodal stimuli (Circle-Tone), non-arbitrarily associated multimodal 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 unimodal to multimodal stimuli, we found several regions that were more active to specific multimodal stimuli than to the constituent unimodal stimuli, using a conjunction definition. Consistent with previous audio-visual studies reporting activity in

lateral occipital cortex (Calvert et al. 1999, 2000, 2001; Bischoff et al. 2007; Stevenson and James 2009), we found multimodal responses in bilateral middle occipital and lingual gyri. Apart from Ellipse-Tone, overlapping clusters of activation for every multimodal condition were found in this region. 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, we found a cluster in the right MFG and clusters bilaterally in FFG. By itself, the fact that regions responding to the specific multimodal 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 (Kang et al. 2006; Szycik et al. 2008; Fairhall and Macaluso 2009), this result was unexpected. The latter point could be partially explained by the fact that most of studies that found activation in the MTG compared congruent and incongruent speech, as opposed to comparing multimodal to unimodal stimuli.

However, a secondary analysis which directly compared the responses of multimodal regions to each of the multimodal conditions revealed that a subset of voxels within bilateral MTG, that were identified as multimodal to Circle-Tone, actually showed a larger response to the multimodal speech conditions. Thus, these areas are quite strongly responsive to the multimodal 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 there was a larger response to Ellipse-Speech than Circle-Speech, which may indicate a modulation of processing based upon the similarity to naturalistic audiovisual speech. Unlike the multimodal 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. A post-hoc examination of the response of this region to the other stimulus categories, including unimodal ones, made it clear that the region was not multimodal using the conjunction approach because it was highly active to any visual stimulus. However, activation was highest during the Ellipse-Speech condition, while activation during Circle-Speech was lowest relative to 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). Thus it may be that while the fusiform is not typically identified as multimodal using standard thresholds, it may indeed be importantly related to audio-visual speech.

Conclusion

Using sets of auditory, visual, and audiovisual stimuli that were matched in terms of their physical properties, this experiment demonstrated differential activation of cortical networks specific to the integration of speech and non-speech stimuli. In particular, the bilateral STG and right MOG exhibited greater activation to all speech conditions than 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 multimodal speech perception. Accordingly, these brain regions may be pivotal in modulating increasingly complex polysensory stimulation.

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

- Alsius A, Navarra J, Campbell R, Soto-Faraco S. Audiovisual integration of speech falters under high attention demands. *Curr Biol*. 2005; 15:839–843. [PubMed: 15886102]
- Arnold P, Hill F. Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. *Brit J Psych*. 2001; 92:339–355.
- Barracough NE, Xiao D, Baker CI, Oram MW, Perrett DI. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *J Cogn Neurosci*. 2005; 17:377–391. [PubMed: 15813999]
- Baumann O, Greenlee MW. Neural correlates of coherent audiovisual motion perception. *Cereb Cortex*. 2007; 17:1433–1443. [PubMed: 16928890]
- Bernstein, LE.; Lu, ZL.; Jiang, JT. Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. Paper presented at the 8th International Multisensory Research Forum; Sydney, AUSTRALIA. 2007, Jul;
- Bischoff M, Walter B, Blecker CR, Morgen K, Vaitl D, Sammer G. Utilizing the ventriloquism-effect to investigate audio-visual binding. *Neuropsychologia*. 2007; 45:578–586. [PubMed: 16620884]
- Boynton GM, Engel SA, Glover GH, Heeger DJ. Linear Systems Analysis of Functional Magnetic Resonance Imaging in Human V1. *J Neurosci*. 1996; 16:4207–4221. [PubMed: 8753882]
- Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, David AS. Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport*. 1999; 10:2619–2623. [PubMed: 10574380]
- Calvert GA, Campbell R, Brammer J. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr Biol*. 2000; 10:649–657. [PubMed: 10837246]
- Calvert GA, Hansen PC, Iversen SD, Brammer MJ. Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage*. 2001; 14:427–438. [PubMed: 11467916]
- Capek CM, MacSweeney M, Woll B, Waters D, McGuire PK, David AS, et al. Cortical circuits for silent speechreading in deaf and hearing people. *Neuropsychologia*. 2008; 46:1233–1241. [PubMed: 18249420]
- Cotton JC. Normal 'visual hearing'. *Science*. 1935; 82:592–593. [PubMed: 17741539]
- Degerman A, Rinne T, Pekkola J, Autti T, Jääskeläinen IP, Sams M, et al. Human brain activity associated with audiovisual perception and attention. *Neuroimage*. 2007; 34:1683–1691. [PubMed: 17204433]
- Dhamala M, Assisi CG, Jirsa VK, Steinberg FL, Kelso JAS. Multisensory integration for timing engages different brain networks. *Neuroimage*. 2007; 34:764–773. [PubMed: 17098445]
- Dixon NF, Spitz LT. The detection of audiovisual desynchrony. *Perception*. 1980; 9:719–721. [PubMed: 7220244]
- Eckert MA, Kamdar NV, Chang CE, Beckmann CF, Greicius MD, Menon V. A cross-modal system linking primary auditory and visual cortices: Evidence from intrinsic fMRI connectivity analysis. *Hum Brain Map*. 2008; 29:848–857.
- Fairhall SL, Macaluso E. Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *Eur J Neuroscience*. 2009; 29:1247–1257.

- Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn Res Med*. 1995; 33:636–647.
- Frens MA, Van Opstal AJ, Van der Willigen RF. Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & Psychophysics*. 1995; 57:802–816. [PubMed: 7651805]
- Goebel R, Esposito F, Formisano E. 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. *Hum Brain Map*. 2006; 27:392–401.
- Grant KW, Ardell LH, Kuhl PK, Sparks DW. The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *J Acoustical Soc Am*. 1985; 77:671–676.
- Grant KW, Seitz PF. The use of visible speech cues for improving auditory detection of spoken sentences. *J Acoustical Soc of Am*. 2000; 108:1197–1208.
- Grant KW, van Wassenhove V, Poeppel D. Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Speech Comm*. 2004; 44:43–53.
- Hocking J, Price CJ. The role of the posterior superior temporal sulcus in audiovisual processing. *Cereb Cortex*. 2008; 18:2439–2449. [PubMed: 18281303]
- Hocking J, Price CJ. Dissociating verbal and nonverbal audiovisual object processing. *Brain Lang*. 2009; 108(2):89–96. [PubMed: 19101025]
- Hughes HC, Reuter-Lorenz PA, Nozawa G, Fendrich R. Visual-auditory interactions in sensorimotor processing: saccades versus manual responses. *J Exp Psych - Hum Per Perform*. 1994; 20:131–153.
- Jones JA, Callan DE. Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *Neuroreport*. 2003; 14:1129–1133. [PubMed: 12821795]
- Kamachi M, Hill H, Lander L, Vatikiotis-Bateson E. 'Putting the face to the voice': matching identity across modality. *Curr Biol*. 2003; 13:1709–1714. [PubMed: 14521837]
- Kang E, Lee DS, Kang HJ, Hwang CH, Oh SH, Kim CS, et al. The neural correlates of cross-modal interaction in speech perception during a semantic decision task on sentences: A PET study. *Neuroimage*. 2006; 32:423–431. [PubMed: 16644239]
- Kayser C, Petkov CI, Logothetis NK. Visual modulation of neurons in auditory cortex. *Cereb Cortex*. 2008; 18:1560–1574. [PubMed: 18180245]
- King AJ, Palmer AR. Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Exp Brain Res*. 1985; 60:492–500. [PubMed: 4076371]
- Lancaster JL, Woldorff MG, Parsons LM, Liotti M, Freitas CS, Rainey L, Kochunov PV, Nickerson D, Mikiten SA, Fox PT. Automated Talairach Atlas labels for functional brain mapping. *Human Brain Map*. 2000; 10:120–131.
- Laurienti PJ, Perrault TJ, Stanford TR, Wallace MT, Stein BE. On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Exp Brain Res*. 2005; 166:289–297. [PubMed: 15988597]
- Laurienti PJ, Wallace MT, Maldjian JA, Susi CM, Stein BE, Burdette JH. Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices. *Hum Brain Map*. 2003; 19:213–223.
- Martuzzi R, Murray MM, Michel CM, Thiran JP, Maeder PP, Clarke S, et al. Multisensory interactions within human primary cortices revealed by BOLD dynamics. *Cereb Cortex*. 2007; 17:1672–1679. [PubMed: 16968869]
- Massaro DW, Cohen MM. Speech perception in perceivers with hearing loss: synergy of multiple modalities. *J Speech Lang Hearing Res*. 1983; 42:21–41.
- McGrath M, Summerfield Q. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *J Acoustical Soc Am*. 1985; 77:678–685.
- McGurk H, MacDonald J. Hearing lips and seeing voices. *Nature*. 1976; 265:746–748. [PubMed: 1012311]
- Meredith MA, Stein BE. Interactions among converging sensory inputs in the superior colliculus. *Science*. 1983; 221:389–391. [PubMed: 6867718]

- Meredith MA, Stein BE. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J Neurophysiol.* 1986; 56:640–662.
- Miller J. Divided attention: evidence for coactivation with redundant signals. *Cogn Psych.* 1982; 14:247–279.
- Miller LM, D’Esposito M. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J Neurosci.* 2005; 25:5884–5893. [PubMed: 15976077]
- Neely KK. Effect of visual factors on the intelligibility of speech. *J Acoustical Soc Am.* 1956; 28:1275–1277.
- Ojanen V, Mottonen R, Pekkola J, Jaaskelainen IP, Joensuu R, Autti T, et al. Processing of audiovisual speech in Broca’s area. *Neuroimage.* 2005; 25:333–338. [PubMed: 15784412]
- Olson IR, Gatenby JC, Gore JC. A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cogn Brain Res.* 2002; 14:129–138.
- Reisberg, D.; McLean, J.; Goldfield, A. Easy to hear but hard to understand: a speechreading advantage with intact auditory stimuli. In: Dodd, B.; Campbell, R., editors. *Hearing by Eye: The Psychology of Lip-Reading.* London: Lawrence Erlbaum Associates; 1987. p. 97-113.
- Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. *Science.* 1981; 212:947–949. [PubMed: 7233191]
- Robert-Ribes J, Schwartz J-L, Escudier P. A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Rev.* 1995; 9:81–104.
- Robins DL, Hunyadi E, Schultz RT. Superior temporal activation in response to dynamic audiovisual emotional cues. *Brain and Cogn.* 2009; 69:269–278.
- Rosen SM, Fourcin AJ, Moore BCJ. Voice pitch as an aid to lipreading. *Nature.* 1981; 291:150–152. [PubMed: 7231534]
- Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. *Science.* 1995; 270:303–304. [PubMed: 7569981]
- Saito DN, Yoshimura K, Kochiyama T, Okada T, Honda M, Sadato N. Cross-modal binding and activated attentional networks during audio-visual speech integration: a functional MRI study. *Cereb Cortex.* 2005; 15:1750–1760. [PubMed: 15716468]
- Sekiyama K, Kanno I, Miura S, Sugita Y. Auditory-visual speech perception examined by fMRI and PET. *Neurosci Res.* 2003; 47:277–287. [PubMed: 14568109]
- Sestieri C, Di Matteo R, Ferretti A, Del Gratta C, Caulo M, Tartaro A, et al. “What” versus “Where” in the audiovisual domain: An fMRI study. *Neuroimage.* 2006; 33:672–680. [PubMed: 16934491]
- Stanford TR, Stein BE. Superadditivity in multisensory integration: putting the computation in context. *Neuroreport.* 2007; 18:787–792. [PubMed: 17471067]
- Stevenson RA, James TW. Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage.* 2009; 44:1210–1223. [PubMed: 18973818]
- Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. *J Acoustical Soc Am.* 1954; 26:212–215.
- Summerfield Q. Use of visual information for phonetic perception. *Phonetica.* 1979; 36:314–331. [PubMed: 523520]
- Summerfield, Q. Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd, B.; Campbell, R., editors. *Hearing by Eye: The Psychology of Lip-Reading.* London: Lawrence Erlbaum Associates; 1987. p. 3-51.
- Summerfield Q, McGrath M. Detection and resolution of audio-visual incompatibility in the perception of vowels. *J Exp Psych.* 1984; 36A:51–74.
- Szyzik GR, Tausche P, Munte TF. A novel approach to study audiovisual integration in speech perception: Localizer fMRI and sparse sampling. *Brain Res.* 2008; 1220:142–149. [PubMed: 17880929]
- Talairach, J.; Tournoux, P. *Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System: An Approach to Cerebral Imaging.* Thieme; Stuttgart: 1988.
- Watkins S, Shams L, Josephs O, Rees G. Activity in human V1 follows multisensory perception. *Neuroimage.* 2007; 37:572–578. [PubMed: 17604652]

- Watkins S, Shams L, Tanaka S, Haynes JD, Rees G. Sound alters activity in human V1 in association with illusory visual perception. *Neuroimage*. 2006; 31:1247–1256. [PubMed: 16556505]
- Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb Cortex*. 2003; 13:1034–1043. [PubMed: 12967920]

Figure 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 audiovisual stimuli, together with examples of movie stills sampled from the circle and ellipse conditions for the first utterance at three time points.

Figure 2.

Regions showing significance ($p < 0.01$, corrected for multiple comparisons) in a conjunction of contrasts of the multimodal 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.

Figure 3. Regions showing a significant interaction ($p < 0.01$, corrected for multiple comparisons) among multimodal conditions. Regions are color-coded to indicate corresponding plot of mean betas for the region.

Figure 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.

Table 1

Peak Voxels for Conjunction Analysis

Condition	Location	Hemisphere	Talairach Coordinates			
			X	Y	Z	Extent
Circle-Speech	Middle Occipital Gyrus	R	25	-84	-4	20
Circle-Tone	Middle Temporal Gyrus	R	53	-34	8	14
	Superior Temporal Gyrus	R	46	-42	8	15
	Fusiform Gyrus	R	38	-39	-16	8
	Lingual Gyrus	R	25	-78	-4	33
	Lingual Gyrus	L	-23	-78	-4	26
	Middle Temporal Gyrus	L	-58	-35	11	17
Ellipse-Speech	Middle Frontal Gyrus	R	46	9	32	13
	Fusiform Gyrus	R	34	-59	-16	41
	Middle Occipital Gyrus	R	20	-87	-6	36
	Medial Frontal Gyrus	L	-6	45	-8	8
	Lingual Gyrus	L	-14	-87	-8	45
	Fusiform Gyrus	L	-37	-63	-15	21
	Middle Temporal Gyrus	L	-47	-60	7	10

Clusters found to be significant to a conjunction of contrasts of the multimodal condition to respective unimodal conditions ($p < 0.01$, corrected for multiple comparisons using cluster thresholds). Extent of regions in specified in functional voxels (3mm^3). Anatomical labels were derived from the Talairach.org database (Lancaster, Woldorff, Parsons, Liotti, Freitas, Rainey L, Kochunov, Nickerson, Mikiten, Fox, 2000).

Table 2

Peak Voxels for Interaction Effects among Conjunction Regions

Talairach Coordinates					
Location	Hemisphere	X	Y	Z	Extent
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 (3mm^3). Anatomical labels were derived from the Talairach.org database (Lancaster, et al., 2000).