Reading Speech from Still and Moving Faces: The Neural Substrates of Visible Speech

Gemma A. Calvert[1] and Ruth Campbell[2]

Abstract

■ Speech is perceived both by ear and by eye. Unlike heard speech, some seen speech gestures can be captured in stilled image sequences. Previous studies have shown that in hearing people, natural time-varying silent seen speech can access the auditory cortex (left superior temporal regions). Using functional magnetic resonance imaging (fMRI), the present study explored the extent to which this circuitry was activated when seen speech was deprived of its time-varying characteristics.

In the scanner, hearing participants were instructed to look for a prespecified visible speech target sequence (''voo'' or ''ahv'') among other monosyllables. In one condition, the image sequence comprised a series of stilled key frames showing apical gestures (e.g., separate frames for ''v'' and ''oo'' [from the target] or ''ee'' and ''m'' [i.e., from nontarget syllables]). In the other condition, natural speech movement of the same overall segment duration was seen.

In contrast to a baseline condition in which the letter ''V'' was superimposed on a resting face, stilled speech face images generated activation in posterior cortical regions associated with the perception of biological movement, despite the lack of apparent movement in the speech image sequence. Activation was also detected in traditional speech-processing regions including the left inferior frontal (Broca's) area, left superior temporal sulcus (STS), and left supramarginal gyrus (the dorsal aspect of Wernicke's area). Stilled speech sequences also generated activation in the ventral premotor cortex and anterior inferior parietal sulcus bilaterally.

Moving faces generated significantly greater cortical activation than stilled face sequences, and in similar regions. However, a number of differences between stilled and moving speech were also observed. In the visual cortex, stilled faces generated relatively more activation in primary visual regions (V1/V2), while visual movement areas (V5/MT+) were activated to a greater extent by moving faces. Cortical regions activated more by naturally moving speaking faces included the auditory cortex (Brodmann's Areas 41/42; lateral parts of Heschl's gyrus) and the left STS and inferior frontal gyrus.

Seen speech with normal time-varying characteristics appears to have preferential access to ''purely'' auditory processing regions specialized for language, possibly via acquired dynamic audiovisual integration mechanisms in STS. When seen speech lacks natural time-varying characteristics, access to speech-processing systems in the left temporal lobe may be achieved predominantly via action-based speech representations, realized in the ventral premotor cortex. ■

INTRODUCTION

Speechreading is the ability to understand a spoken message by watching the speech actions of a talker. It has traditionally been seen as a topic of interest to clinical researchers interested in the implications of hearing loss, deafness (Jeffers & Barley, 1971), or hearing in noise (Sumby & Pollack, 1954), but is increasingly seen to have implications for understanding the mechanisms of speech and language processing more generally (Liberman & Whalen, 2000; Green, 1998; Dodd & Burnham, 1988). This is because all people who use speech are sensitive to its visual qualities, despite the fact that individual speechreading abilities can vary widely. For example, audiovisual speech perception is reliably better than the perception of speech that is simply heard (Sumby & Pollack, 1954), even when auditory speech is perfectly clear (Reisberg, McLean, & Goldfield, 1987).

Behavioral studies have shown that hearing infants are sensitive to audiovisual speech synchronization (Dodd, 1979) and to the fit of the seen and heard speech characteristics, including the discrimination of the vowel that is uttered, the identity of the speaker, and the type of utterance produced (Lewkowicz, 1996, 1998; Burnham, 1993; Kuhl & Meltzoff, 1982). Susceptibility to audiovisual speech illusions, whereby a dubbed utterance, such as seen ''ga'' with heard ''ba'' is perceived as ''da'' (McGurk & MacDonald, 1976), has also been demonstrated in infants (Rosenblum, Schmuckler, & Johnson, 1997). Studies in adults (Massaro, 1998) have shown that these audiovisual speech illusions are not an isolated phenomenon, but evidence of systematic integration of seen and heard speech in normal speech processing (e.g., Massaro, 1999). Indeed, adult audiovisual speech processing is

[1] University of Oxford, [2] University College London

sensitive to native heard language structure (Sekiyama, 1997; Sekiyama & Tohkura, 1993; Werker, Frost, & McGurk, 1992) and to a range of visible perceived talker characteristics (Green, Kuhl, Meltzoff, & Stevens, 1991). Only models of speech perception that go beyond auditory processing, and implicate "supramodal" or "amodal" procedures can accommodate such findings (see Green et al., 1991; Summerfield, 1987, 1992; Fowler & Rosenblum, 1991). If speechreading is intrinsic to an understanding of speech processing, one important question that arises is, which of its visual stimulus dimensions or properties are utilized by the speech-processing system? Here, two contrasting possibilities are outlined, which have implications for understanding the cortical bases of speechreading and its relation to heard speech.

Time-Varying Information in Seen Speech

The pattern of movement made by a face can be captured by a point-light display using sparse (8–30) illuminated points on the facial surface, including the cheeks, lips, chin, and nose. Under these conditions, face features such as the mouth, lips, and tongue cannot be reliably identified. That is, the configural image properties of the face and mouth are impoverished or absent. However, when the speaker's face moves in speech such that the illuminated points follow the appropriate trajectories, these point-light displays can affect the accuracy with which auditory speech tokens are identified (Rosenblum, Johnson, & Saldaña, 1996; Rosenblum & Saldaña, 1996). An explanation of this effect may lie in the fact that the actions of the articulators have both visible and audible consequences that are likely to be highly correlated with each other because of their common source properties in the vocalizations of the talker.

One such property is the timing of changes in vocalization—the dynamic properties of speech are visible as well as audible in terms of their time-varying patterns (Munhall & Vatikiotis-Bateson, 1998). For example, increases in speech sound amplitude can be accompanied by visible indicators of change in the disposition of the visible articulators—such as the speed and acceleration of mouth opening. The correlation between some auditory and visual dynamic patterns in utterances can be very striking indeed. For example, for one talker, the fundamental frequency (F_O) components of the speech stream over a sentence-long utterance can be predicted with >90% accuracy simply by tracking the position of the talker's moving head (Yehia, Rubin, & Vatikiotis-Bateson, 1998).

Given the high correlation between time-varying characteristics of the auditory and visual components of speech, audiovisual speech may be comprehensible even when information from one or other channel (vision or audition) is degraded. In the limiting case, when both the visual and auditory streams have each been degraded to a level at which speech cannot be identified, the audiovisual stream may yet be understandable because of redundancy in the dynamic patterning of speech across the two modalities. Several studies show an influence of vision on auditory speech even when the individual speech events cannot be discriminated by eye. For instance, Jordan and Sergeant (2000) showed that vision could affect the report of auditory syllables at viewing distances too great for the visual syllable to be identified, yet sufficiently close for the seen action to be perceived as plausibly congruent with the heard syllable. Grant and Seitz (2000) have shown that correlated information from the face in motion improves detection of noisy auditory messages even though neither visual nor auditory segments could be identified reliably on their own. Such demonstrations suggest that the visible dynamic signature of a spoken utterance is informative when segmental properties of speech within either the visible or auditory speech stream are not fully accessible. Its utility lies in the redundancy of information perceived from the talking head, in particular in the common dynamic properties of the utterance, whether seen or heard.

A dynamic systems (time-varying) account of speechreading, therefore, does not require the perceiver to identify a particular image component of the speaking face. Even if the form of the facial image is underspecified, vision can nevertheless improve speech processing. However, a completely contrary case can also be made for visible speech processing—that good image processing in the absence of well-specified time-varying information is an important feature of multimodal speech.

Configural (Image-Based) Considerations in Speechreading

Although the time-varying regularities in seen and heard speech are used in speech processing, human sensitivity to audiovisual synchronization is often quite poor. Imperfect time-streaming of videoclips in digitized audiovisual speech segments, where synchrony of the seen and heard message is poorly preserved, may not be noticeable. Campbell and Dodd (1980) reported an advantage to audiovisual speech processing in noise even when vision and audition were desynchronized by 1.5 sec. One reason may be that when visual information is relatively well specified at the image level it can be utilized by the speech-processing system, despite poor time-varying correlations with heard speech properties. While heard speech cannot be identified from a non-time-varying depiction of speech information, for example, from a picture of a speech spectrogram, this is not the case for seen speech. The stilled face image can be speechread (see Figure 1)

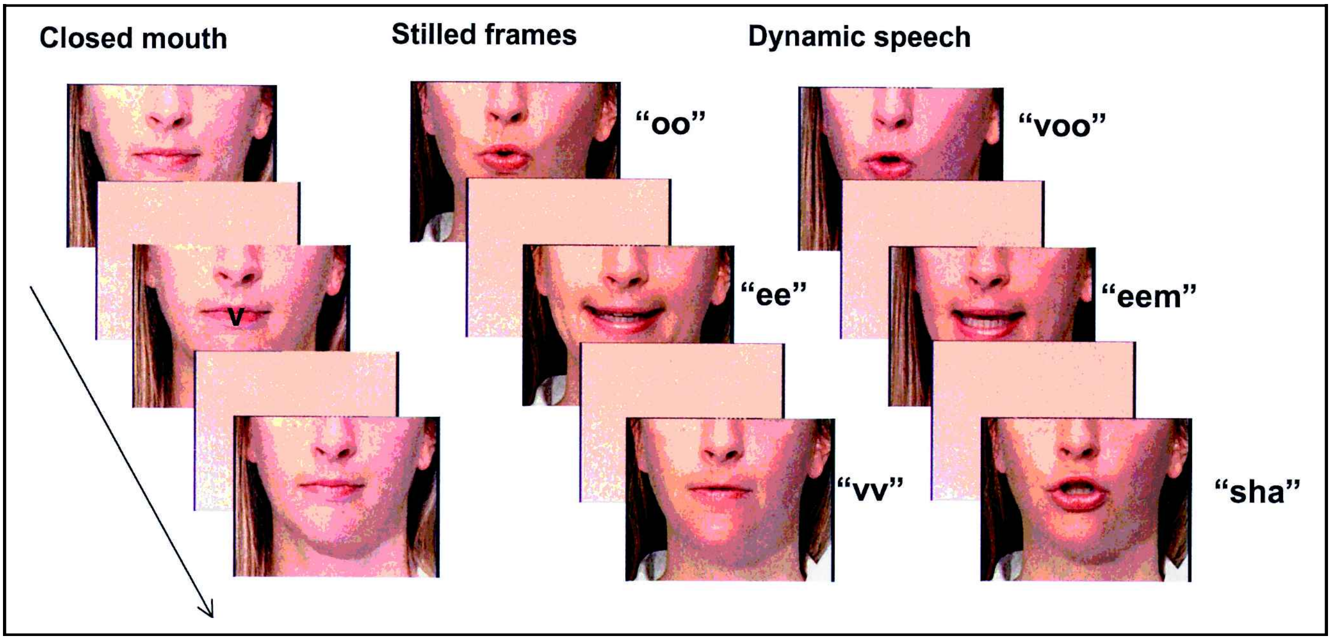

Figure 1. Examples of the stimuli presented in the closed mouth (control), stilled speech frame, and dynamic speech conditions. All facial images were interleaved between luminance-matched skin-tone frames to prevent apparent motion and flicker artefacts, and presented in 30-sec blocks with 10 trials per block.

and experimental studies have made use of this to explore the functional separability of reading speech, identity and emotion from the facial image (e.g., Schweinberger & Soukup, 1998; Campbell, Brooks, de Hann, & Roberts, 1996).

The ability to distinguish stilled images of visible vowel shapes ("is it 'ee' or 'oo'?"), or labials from velars ("is it 'b' or 'k'?") is relatively easy. For apical segments (i.e., the point at which the utterance has the most characteristic and distinctive phonological structure), visible image properties can often be sufficient to distinguish speech sounds, if not to identify all of them. Facial images showing mouth shape, lips, tongue, and teeth position offer potential information about the filter state of the vocal tract (Summerfield, 1992). Moreover, when the image properties of the face are disturbed while time-varying properties are maintained, the influence of vision on audition is reduced. Changes such as reversal of contrast polarity and inverting the orientation of the face (Jordan & Bevan, 1997; Jordan, McCotter, & Thomas, 2000; Massaro & Cohen, 1996) reduce susceptibility to audiovisual speech illusions.

Direct evidence that stilled speech images affect auditory speech perception comes from demonstrations that such images can even generate McGurk effects when combined with heard speech. For example, Whalen, Irwin, and Fowler (in press) have found that dubbing a stilled image of a consonant that is not congruent with one that is heard can generate reports of an "illusory" consonant in consonant–vowel utterances (monosyllables) in some perceivers. Cathiard, Tiberghien, and Arby (1992) and Cathiard and Tiberghien (1994) showed that synchronizing a seen "oo" face image to a heard "ee" speech sound generated the perception of /y/ (as in the French "lune") in French-speaking viewers.

The studies reviewed suggest that "both" configural (stilled) and time-varying (naturally moving) face actions play a role in the perception of speech. Configural information, available from the stilled image, may be especially useful in delivering specific face-action patterns that suggest a particular phonetic gesture or type of articulation. Time-varying information may be especially useful in tracking a range of commonalities across heard and seen speech, which are reflected, redundantly, in the common dynamic structure of both. The question then arises: If both stilled speech images, and depictions of moving faces in action are each readily incorporated into speech processing, do they make use of different cortical circuitry—or is seen speech processed in an identical manner whether it is delivered by time-varying or configural means?

Cortical Considerations

Neuropsychological studies suggest that acquired impairments in the perception of moving and of stilled images show dissociable effects on speechreading. Campbell, Zihl, Massaro, Munhall, and Cohen (1997) report that the movement-blind patient, LM, could identify stilled but not moving visible speech patterns. She could not speechread naturally moving mouths, showed no audiovisual illusion sensitivity, and could not interpret point-light speech displays. A contrasting pattern was obtained in the visual agnosic patient HJA

(Campbell, 1992), who was unable to ''see'' stilled images of face actions, but showed normal sensitivity to natural facial movement in his susceptibility to audiovisual speech illusions. In HJA, damage was bilateral and confined to the primary visual cortex (V1/2), while in LM it was specific to region MT/MST, bilaterally. Thus, their functional deficits in reading speech from faces reflect damage to specific visual input systems. Both these patients appeared to be able to process faces under appropriate viewing conditions.

The cortical substrates of face-image processing have now been widely investigated using neuroimaging techniques (see Allison, Puce, & McCarthy, 2000; Haxby, Hoffman, & Gobbini, 2000; Kanwisher & Moscovitch, 2000 for recent reviews and discussions). So, too, have those for visible speech (Olson, Gatenby, & Gore, 2002; Bernstein et al., 2002; Campbell et al., 2001; Surguladze et al., 2001; Callan, Callan, Kroos, & Vatikiotis-Bateson, 2000; Ludman et al., 2000; MacSweeney et al., 2000, 2001; Levänen, 1999; Calvert et al., 1997, 1999; Calvert, Campbell, & Brammer, 2000; Sams et al., 1991; Sams & Levänen, 1996).

The processing of face images utilizes specialized inferotemporal regions of the right hemisphere, especially the middle part of the fusiform gyrus—the face fusiform area (FFA; Kanwisher, McDermott, & Chun, 1997). How faces are processed beyond FFA appears to depend on functional task requirements: A wide range of (generally) right hemisphere localized structures have been implicated, depending on task. Regions implicated include many temporal regions, including the temporal pole (facial identity), medial temporal regions including the hippocampus (face memory tasks), superior parietal, and frontal regions (gaze and some facial expression tasks: see Haxby et al., 2000 for a review).

In contrast to most tasks that require speeded processing of the facial image, which tend to show a right hemisphere advantage, silent stilled-speech image matching can show a left hemisphere bias (Campbell, De Gelder, & De Haan, 1996). Interestingly, while several neuroimaging studies of face processing report activation in middle and superior temporal regions, the processing of the stilled facial image does not appear to activate these bilateral areas consistently, suggesting that activation here may be task, rather than stimulus specific.

STS Specialization and Reading Speech

Bonda, Petrides, Ostry, and Evans (1996) and Howard et al. (1996) were the first to show that the perception of a dynamic array of point-lights representing a person in movement specifically activates a focal site on the superior temporal gyrus, along the ventral bank of the posterior superior temporal sulcus (STS). Recent studies uphold the conclusion that the STS is activated specifically by biological motion (Grossman et al., 2000). Activation of the STS has been reported during the perception of eye and mouth movements (Puce, Allison, Bentin, Gore, & McCarthy, 1998; see Allison et al., 2000 for a review), and when viewing nonspeech facial movement (e.g., gurning movements: Campbell et al., 2001; Calvert et al., 1997).

''All'' studies of natural speechreading and audiovisual speech (MacSweeney et al., 2002; Calvert et al., 1997, 1999, 2000) show consistent and extensive activation of STS in hearing people. Given the crucial sensitivity of STS to the dynamic patterning of seen biological events, including facial actions, as well as to its role in audiovisual speech processing, it seems plausible that this is one region that will show differential sensitivity to moving speech. We have pointed out that time-varying structure is correlated across seen and heard speech and enhances speech processing. Calvert et al. (1999, 2000) have proposed that heteromodal regions within the STS may perform a specific cross-modal binding function. For audiovisual speech that is appropriately synchronized, the profile of the activation in STS has been shown to correlate with enhanced neuronal activity in sensory-specific visual (V5/MT) and auditory (A1/2) cortices (Calvert et al., 2000; Sams et al., 1991). These cross-modal gains may be mediated via back projections from putative binding sites in STS (Calvert et al., 1999). Since natural speech that is heard and seen shares a unitary dynamic structure, this may account for the ability of dynamically structured seen speech to activate auditory language processing regions, including early auditory processing areas (Heschl's gyrus and surrounding belt region) in the left hemisphere, in the absence of heard speech.

The functional magnetic resonance imaging (fMRI) study reported here was designed so that it could be performed using stilled-image sequences as well as speech seen in its natural dynamic state. This was the detection of a highly speech-readable silently spoken target syllable, /a:v/ (''ahv'') or /vu:/ (''voo'') among a list of other syllables e.g., ''boo,'' ''eem,'' ''shah'' (/bu:/, /Im/, /sha:/), all of which are visibly distinctive and not confusable with the target. We presented the material in two ways: as natural movement sequences and as sequences of stilled frames using just the apical gestures (see Figure 1). The production of this sequence was tightly controlled and tested to ensure that no apparent movement was visible between frame shifts (see Methods).

Whole brain fMRI scans at 3T were performed, and all the participants were investigated under three conditions interleaved within a block design: Watching a face at rest with a V superimposed for an unpredictable 1-sec duration on the lip region (control condition); watching a series of stilled apical gesture speech sequences without movement (stilled speech); and watching a series of naturally moving speech sequences (moving speech). In each condition, the targets: ''ahv'' and ''voo'' (moving condition), a stilled image of the mouth in the midst of pronouncing "V" (stilled condition), or the letter "V" on a closed mouth image (control condition) occurred unpredictably twice in each block of 10 trials. Subjects were required to "press the button whenever they saw a 'v'—as a gesture or as a letter." They made a button press when targets were detected.

In addition to distinctive activation in posterior (visual) regions between dynamic and stilled speech, reflecting their differing reliance on movement (V5/MT and V1/V2), we hypothesized that stilled speech may not access all the regions implicated in previous speechreading studies. In particular, portions of STS and STG activated by natural speech may show relatively reduced activation. These regions may be especially sensitive to the common dynamic structure of seen and heard speech.

RESULTS

Behavioral Results

Mean accuracy of response was 98% for the written target, 66.5% (*SD* = 14.3) for the stilled condition; 68.1% (*SD* = 13.3) for the moving condition. Following an arcsine transformation of the data to normalize ceiling effects, one-way analysis of variance showed that the written target condition was performed more accurately ($p < .01$) than either of the speech target conditions, which did not differ. This is unsurprising since the "only" event in the control condition was

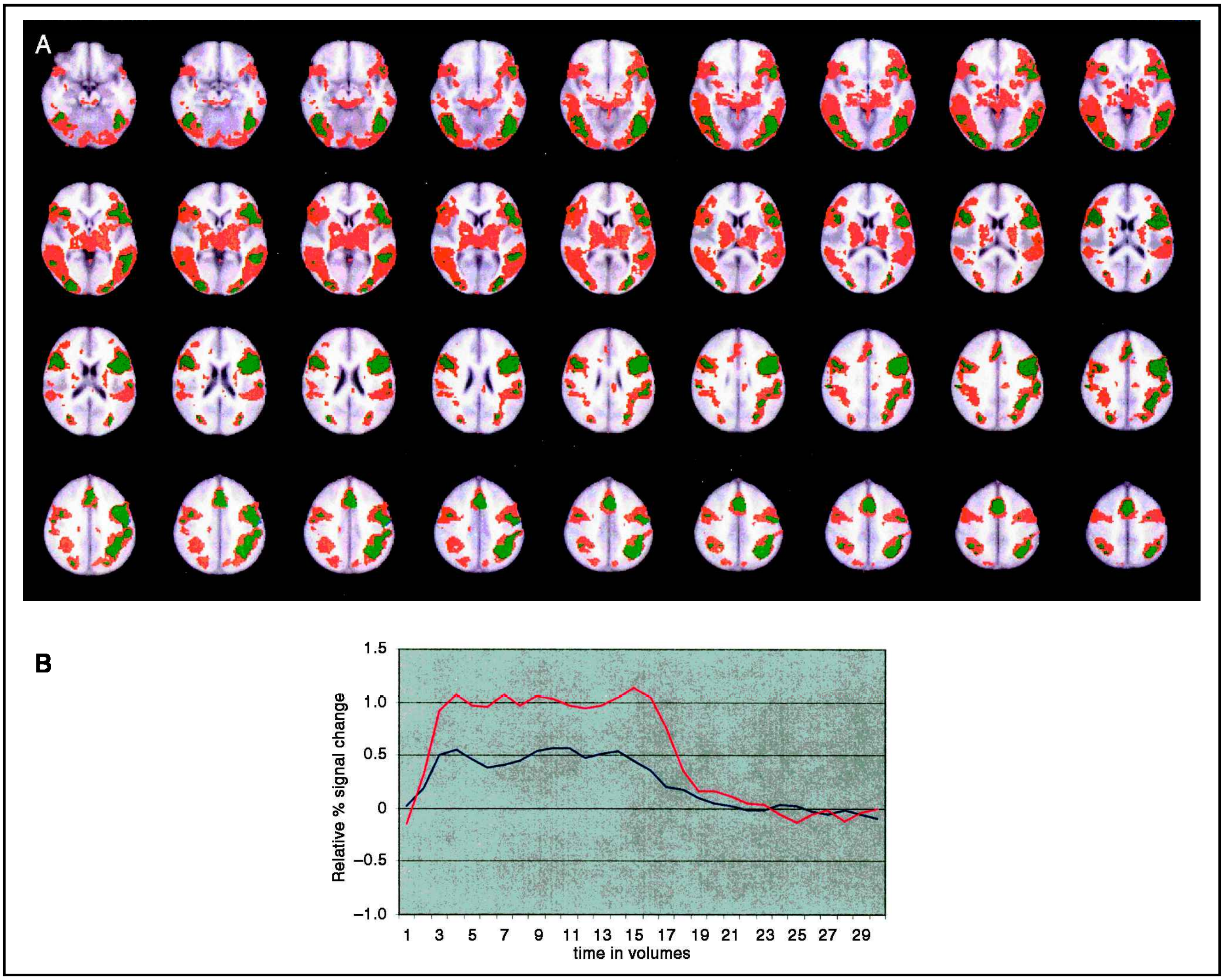

Figure 2. (A) The figure shows each experimental condition compared with rest. Voxels colored red were activated by dynamic speech alone. Voxels activated solely by stilled speech were colored blue, and green voxels represent the areas activated "both" by stilled and moving speech. As can be seen from the image, there are no regions activated by stilled speech (i.e., blue voxels) that are not also activated by moving speech. The images are shown in radiological convention so that the left of the each brain slice reflects the right hemisphere. (B) Summarized mean image intensity across the entire network of areas activated in stilled and moving speech conditions compared to the control condition. Activation to moving speech is nearly double the amplitude of the stilled speech condition. Average time-course for the group: pink = dynamic images (on) versus control (off), blue = stilled images of speech (on) versus control (off) averaged across the whole scanning period.

the occurrence of the target letter, while in the speech conditions the target syllable was embedded in other syllables.

Imaging Results

Stilled Speech—Control Condition

Stilled speech activated a network of brain areas, several components of which have previously been implicated in reading and lipreading. These included bilateral regions of inferior and middle frontal gyri (BA 45/45; 44/9), which were stronger and more extensive in the left hemisphere; as well as areas within the medial frontal (BA 6/8) and precentral (BA 4/6) gyri bilaterally. In the visual cortex, bilateral activation was observed in the fusiform gyri (BA 19/37) extending superiorly into the inferior and middle occipital gyri (BA 18/19) and superior-anteriorly towards the occipito-temporal

Table 1. Brain Areas Activated in the Group-Averaged Contrasts of the Two Experimental Conditions Versus The Closed Mouth Baseline Condition

		Talairach Coordinates					*Talairach Coordinates*				
		Side	*x*	*y*	*z*		*Side*	*x*	*y*	*z*	
Anatomical Region	*BA*	*Moving Versus Closed Mouth*				*Z Score*	*Stilled Versus Closed Mouth*				*Z Score*
Visual cortex											
Fusiform gyrus	19/37	L	−44	−66	−10	20.2	L	−44	−54	−14	9.1
		R	48	−64	−10	20.1	R	46	−58	−12	7.1
Inferior occipital gyrus	18	L	−24	−96	−2	13.6	L	−26	−94	0	7.2
		R	28	−88	0	15.8	R	24	−90	−2	6.6
Middle occipital gyrus	19	L	−28	−74	18	6.7	L	−30	−82	6	7.5
		R	30	−74	18	6.8	R	26	−88	12	6.4
Occipito-temporal junction [V5/MT]	19/37	L	−44	−68	−6	24.8	L	−48	−64	−8	11.5
		R	46	−62	−4	25.5	R	54	−52	−4	8.2
Temporal cortex											
Heschl's gyrus	41	L	−54	−14	10	10.5					
Superior temporal gyrus	42/22	L	−50	−38	12	11.4					
		R	58	−34	12	6.4					
Superior temporal sulcus	22/21	L	−48	−46	4	21.0	L	−50	−42	4	6.5
		R	54	−50	4	20.2					
Middle temporal gyrus	21	L	−56	−44	2	19.2	L	−54	−46	−2	6.4
		R	52	−54	2	20.2	R	54	−52	−4	8.2
Parietal cortex											
Inferior parietal lobe	40/39	L	−32	−52	40	12.9	L	−32	−58	46	9.3
		R	34	−48	40	6.7	R	42	−46	46	6.5
Superior parietal lobe	7	L	−32	−58	46	12.7	L	−32	−58	54	6.5
		R	28	−56	50	6.8	R	40	−48	52	6.3
Frontal cortex											
Inferior frontal gyrus	44/45	L	−48	26	6	13.4	L	−46	28	6	8.2
		R	44	12	22	13.5	R	38	26	−2	6.2
Middle frontal gyrus	44/9	L	−40	10	26	20.4	L	−42	10	28	12.2
		R	44	14	24	11.1	R	44	12	20	9.2
Medial frontal gyrus	6/8/9	L	−2	22	48	13.2	L	4	24	46	6.1
		R	4	6	56	14.0	R	2	22	46	8.9
Precentral gyrus/sulcus	4/6/8	L	−44	−4	42	13.8	L	−52	−8	42	7.8
		R	46	2	44	13.4	R	46	0	44	6.5

Table 1. (*continued*)

		Talairach Coordinates					*Talairach Coordinates*				
		Side	*x*	*y*	*z*		*Side*	*x*	*y*	*z*	
Anatomical Region	*BA*	*Moving Versus Closed Mouth*				*Z Score*	*Stilled Versus Closed Mouth*				*Z Score*
Other regions											
Cerebellum		L	−38	−70	−28	11.3	L	−40	−70	−26	6.5
		R	8	−80	−32	12.2	R	34	−54	−28	7.3
Thalamus		L	−12	−14	2	9.7					
		R	−16	−18	2	6.9					

junction (BA 19/37). The *x*, *y*, and *z* plane coordinates of visual motion areas have been putatively identified on the basis of a meta-analysis of relevant functional imaging studies (Hasnain, Fox, & Woldorff, 1998). Activations in the visual cortex extended into these regions despite the lack of real or apparent movement in the stilled speech stimuli.

Activations were also observed in the parietal lobe extending upwards from the left postcentral gyrus (BA 40/43) into the inferior parietal lobule and supramarginal-angular gyrus (BA 40/39) border. The peak of these activations was located in the left inferior parietal sulcus. Activation also extended superiorly into the superior parietal lobules (BA 7). In the temporal lobe, clusters of activation were localized to the middle temporal gyri (BA 21) and in the left hemisphere, activation in this area extended superiorly into the fundus of the STS (x = −50; y = −44; z = 4). Finally, activation was also observed in the cerebellar hemispheres. Within this network of brain areas, the strongest responding areas (i.e., those exhibiting the highest statistical power) were the left middle frontal gyrus (BA 44/9), the left inferior parietal sulcus (BA 40), and the left occipito-temporal junction (BA 19/37).

Taken together, this pattern suggests that the task of identifying a silent, stilled, spoken syllable activates frontoparietal systems and occipito-temporo-parietal systems that extend beyond primary visual cortex. Despite the control contrast with a face at rest, the experimental condition also elicited greater activation in and around the face fusiform region (FFA). Activation in the supramarginal gyrus, the middle temporal (including STS) and inferior frontal regions all showed stronger activation on the left than the right, implicating traditional speech-processing circuitry.

Moving Speech—Control Condition

In this contrast, moving speech was found to activate all the brain areas stimulated in the stilled speech versus control condition, as well as some additional sites (Figure 2A). However, activation in the moving speech condition was more extensive and of greater magnitude. This was particularly marked in V5/MT and adjacent inferior occipito-temporal regions consistent with the presence of dynamic stimuli. Figure 2B shows the overall amplitude difference in the block-averaged BOLD time course between the moving and stilled conditions (each contrasted against the control condition).

Unlike stilled speech, moving speech (when contrasted against the control condition) activated large swathes of cortex in the lateral superior temporal region. Activation in this area extended into the lateral tip of Heschl's gyrus (A2), in line with previous findings that activation by natural silent speech includes specialized auditory regions (MacSweeney et al., 2000; Calvert et al., 1997, 2000). Moving speech also activated the ventral lateral nuclei of the thalamus.

Moving-Control and Stilled-Control: Similarities and Differences

When contrasted against our control resting face condition, stilled speech activated a subset of the areas activated by normal dynamic speech items (as shown in Table 1). Differences in the relative strength of activation within this common network were also apparent. While the strongest activation in the stilled speech condition was located in the ventral premotor cortex (BA 44/9), in the moving speech condition it was identified in or near visual motion cortex (BA 19/37). These differences in the relative level of activation between the two conditions may suggest a greater influence of top-down versus bottom-up mechanisms in the processing of stilled and moving images of speech respectively.

Effects of Movement in Speech: Moving > Stilled Speech and Stilled > Moving Speech

The pattern of activation produced by both speech conditions included traditional language processing sites comprising both inferior frontal and middle and superior temporal regions of the left hemisphere. However, movement affected the patterns differentially (Table 2). Consistent with their specific stimulus characteristics, moving speech generated greater activation in visual motion areas (V5/MT) extending superiorly and anteriorly into middle and superior temporal gyri (BA 21/22). Other areas activated more by moving than stilled speech include the ventral (bilaterally) and dorsal (left) sectors of the inferior frontal gyrus (BA 47 and 44), ventral premotor cortex bilaterally (BA 9/6), medial dorsal frontal gyrus (BA 6), and the left intraparietal sulcus (BA 40) (Figure 3).

For the stilled versus moving speech contrast, significant clusters were not found when created with the ($Z > 4.7$) cluster-creation threshold, tested at ($p < .01$). However, when the cluster-creation threshold was

Table 2. Brain Areas Activated in the Group-Averaged Contrasts Between Moving and Stilled Speech

		Talairach Coordinates					*Talairach Coordinates*				
		Side	*x*	*y*	*z*		*Side*	*x*	*y*	*z*	
Anatomical Region	*BA*	*Moving Versus > Stilled Speech*				*Z Score*	*Stilled speech > Moving Speech*				*Z Score*
Visual cortex											
Lingual gyrus	19/37						L	−12	−66	2	3.4
							R	8	−70	2	3.3
Inferior occipital gyrus	18	L	−30	−92	−4	6.4					
		R	24	−90	−4	7.4					
Middle occipital gyrus	19						L	−40	−84	26	3.6
Occipito-temporal junction [V5/MT]	19/37	L	−42	−68	−4	14.8					
		R	44	−62	−4	19.3					
Temporal cortex											
Superior temporal gyrus	42/22	L	−48	−34	20	7.6					
		R	54	−36	16	8.0					
Superior temporal sulcus	22/21	L	−48	−50	4	13.2					
		R	58	−30	6	7.5					
Middle temporal gyrus	21	L	−54	−52	0	14.7					
		R	54	−48	0	11.2					
Parietal cortex											
Inferior parietal lobule	39/40	L	−50	−38	26	6.6	L	−44	−50	22	3.4
							R	−48	−54	26	2.8
Superior parietal lobe	7	L	−32	−48	56	6.0					
Frontal cortex											
Inferior frontal gyrus	44/45	L	−34	24	−6	5.8					
		R	40	28	−10	6.1					
Middle frontal gyrus	44/9	L	−40	10	26	8.3					
Medial frontal gyrus	6	M	−2	−2	58	6.5					
Precentral gyrus/sulcus	4/6	L	−36	−2	52	6.4					
Superior frontal gyrus	10/11						L	−10	52	−12	3.8
							R	12	54	−8	3.8
Other regions											
Cerebellum		L	−14	−80	−32	6.7					
		R	10	−80	−26	7.5					
Thalamus		L	−12	−14	2	6.0					
		R	10	−14	2	6.8					

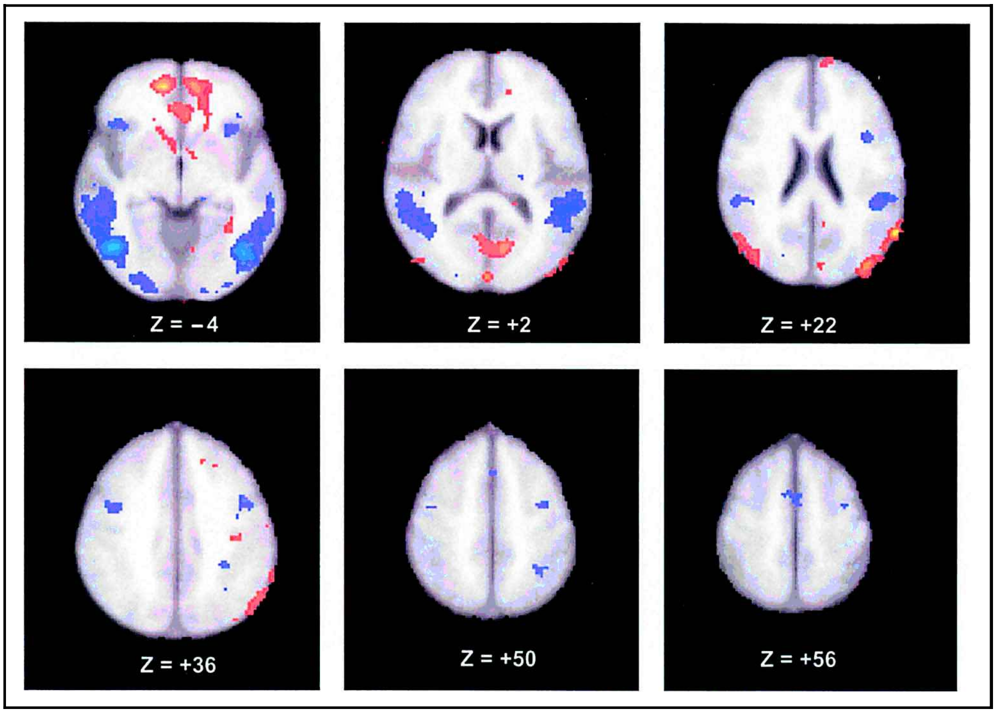

Figure 3. Axial slices at various levels along the z axis exhibiting peak differential activations between the moving—stilled conditions (blue) and stilled—moving conditions (red). Activations are shown in radiological convention. Stilled speech generated greater activation in the orbitofrontal (top left), lingual gyrus (top middle) and superior parietal lobules (top right and bottom left). Moving speech generated greater activation in occipitotemporal and inferior frontal areas (top left) extending into superior temporal cortex (top middle). Additional activation was also detected in inferior parietal sulci and the ventral premotor cortex bilaterally (bottom left and middle) and in the supplementary motor area (bottom right).

dropped to 1.8 (given that "broader signals are best detected by lower thresholds" [Friston, Worsley, Frackowiak, Mazziotta, & Evans, 1994]), several clusters were found that still passed the final significance test ($p < .01$). The largest of these was identified in the lingual gyrus (V1–V2). Weaker responding clusters were also found in the orbitofrontal cortex (BA 10/11) and inferior parietal lobules incorporating foci within the angular and supramarginal gyri (BA 39/40) bilaterally.

DISCUSSION

Seen speech can affect the processing of heard speech—even when it lacks dynamic structure. In this task, of monosyllable target spotting, and despite the differences in "naturalness" of the material, stilled and moving speech sequences were processed equally efficiently. But were the cortical correlates of the two experimental tasks the same? Based on previous findings, and following the argument that the STS may be especially tuned to natural speech in both its visual and auditory aspects, we had predicted that naturally moving speech should have preferential access to STS and the auditory cortex, as well as to visual regions specialized for motion processing.

As expected, activation by natural dynamically structured speech was extensive in posterior regions specialized for visual movement processing (V5/MT). Activation in these areas extended bilaterally into middle and superior temporal regions, including the STS, and rostrally into superior parts of the superior temporal gyrus (including the lateral tip of Heschl's gyrus), that is, into the auditory cortex, replicating previous findings (Olson et al., 2002; Bernstein et al., 2002; Campbell et al., 2001; Surguladze et al., 2001; Ludman et al., 2000; MacSweeney et al., 2000; Calvert et al., 1997, 2000). All previous studies of speechreading have used lexically structured material. In this study, where syllables, not words, were to be identified, the pattern of activation was not noticeably different with respect to activation in these regions. There was also significant activation in inferior frontal regions, including BA 44/45 (Broca's area), again supporting several studies showing such activation when watching mouth actions (Buccino et al., 2000; Campbell et al., 2001, Experiment 1; MacSweeney et al., 2000).

With respect to language processing regions, stilled speech showed a highly similar pattern to that of moving speech. When subjects viewed static images of speech activation was observed in traditional language processing systems, predominantly in the left hemisphere, including inferior frontal (Broca's area) and lateral temporal regions—the latter in the region of the supramarginal gyrus and STS. The general picture (Figure 2A and B) is that stilled faces generated activation in most of the regions activated by faces that move.

When contrasted with the resting face condition, stilled speech images generated activation in primary visual areas, as predicted, but additionally in posterior occipito-temporal sites sensitive to biological motion, including V5/MT. This adds to the growing evidence that stilled images associated with actions can activate cortical areas sensitive to the perception of visual movement (Hermsdörfer et al., 2001; Kourtzi & Kanwisher, 2000;

Senior et al., 2000). It further suggests that representations of the dynamic trajectories of speech events may be activated by stilled speech images.

The pattern of bilateral (L > R) activation for the contrast between the stilled-face and the rest condition suggests that looking at images of a speaking face may be less right hemisphere lateralized than other face-processing tasks (see also Campbell, De Gelder, et al., 1996), while the demonstration of the activation of language processing regions by seen speech confirms the dissociation between identifying speech from faces and identifying emotions or familiarity from faces. It appears, both from these imaging studies and from patterns of dissociated face reading in patients, that left posterior sites in the territory of the middle cerebral artery (middle/posterior temporal regions) may be required for reading speech from face images, while the corresponding region on the right may be required for reading identity or emotion (Haxby et al., 2000; Campbell, Landis, & Regard, 1986). Further imaging studies are required to pursue these contrasts in face-image processing more systematically.

Regions of peak activation offer further clues concerning the core regions that support or connect regions underpinning the relevant function. In this study, the peak of the STS activations for stilled and moving speech showed quite extensive overlap (Figure 4). This is interesting in the light of prior findings suggesting that stilled and moving mouths can activate different regions of the STS under passive observation conditions that do not require speechreading (see Puce et al., 1998). The requirement in the present study to identify a target speech event, common to both the experimental conditions, may have led to construal of a stilled image in terms of some of its dynamic components.

Although the activation generated by stilled speech represents a subset of the areas stimulated by viewing dynamic images of speech (when both were contrasted against the control condition), the "relative" strength of activations within this common network were nevertheless distinguishable. While moving speech exhibited the strongest activations in visual motion areas, stilled speech stimulated ventral premotor cortex and the intraparietal sulcus more prominently. This finding could offer clues concerning the system utilized to see speech in stilled faces: that is, a second route from visual to auditory language processing regions.

Action-Based Perceptual Processing: A (Secondary) Route to Access Speech Representations?

Prefrontal cortical regions are involved in the planning and preparation of bodily actions (Passingham, 1993), and parietofrontal interactions are critical to their correct realization in space and time (Decety & Grèzes, 1999). The discovery of cell tracts within inferior prefrontal regions that are specialized for the "perception" of actions (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996) has, however, revitalized the proposal that the perception, development and maintenance of representations of actions may require recruitment of specific frontal systems. These regions may therefore hold clues to the development of specific cortical systems for processing conspecific imitation and human language and communication (Rizzolatti & Arbib, 1999).

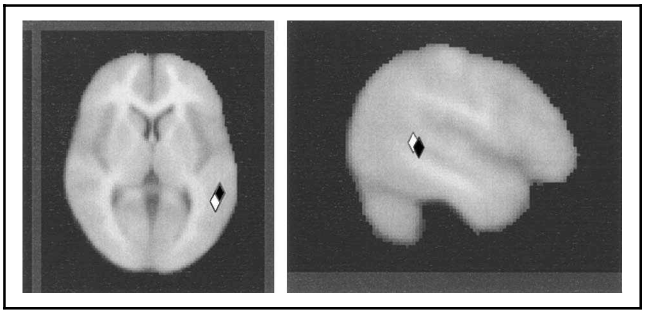

Figure 4. Location of the peak activation in the left STS for dynamic (white diamond) and stilled (black diamond) speech shown in axial ($z = +4$) and sagittal ($x = -48$) orientation in the left and right panels respectively. The foci of these activations are near co-incident.

A number of studies using stilled images of actions have reported activation in prefrontal regions, suggesting that the crucial stimulus cue need not be embodied in the (dynamic) action characteristics of the stimulus, but in its associations with previously encountered actions, that is, that it functions "symbolically." Studies with photographic images of hand gestures report lateral and medial prefrontal activation (BA 44, 45, 8, 9, 6—pre-SMA). These include both imitation and observation studies (Buccino et al., 2001; Hermsdörfer et al., 2001; Iacoboni et al., 1999). Prefrontal regions (BA 8/9) are also implicated in the perception of cartoon story sequences depicting human intentions (Brunet, Sarfati, Hardy-Baylé, & Decety, 2000), as well as in studies focusing on the detection of sequence anomalies in spoken language (Crozier et al., 1999).

The processing of human action sequences thus may recruit prefrontal, inferior frontal and parietal systems independently of the explicit motor requirements of the task (Grèzes & Decety, 2001). Although visuoperceptual tuning of some cells occurs in the inferior prefrontal cortex in the monkey, with analogues in the ventral premotor regions bilaterally, the recruitment of prefrontal-parietal systems more generally is relatively insensitive to stimulus characteristics. Indeed, these systems can be activated by verbal instructions to imagine action sequences (Binkofski et al., 2000).

In the present study, we speculate that stilled images, in contrast to dynamic visible speech, make special use of this circuitry. Following processing in posterior cortex, stilled speech images may access parietofrontal circuits specialized for the observation (and imitation) of human actions. Buccino et al. (2001) have demonstrated distinct localization patterns within the ventral premotor and posterior parietal lobe for identification of photographed actions of mouths, of hands, and of feet. Speech representation systems in the left inferior frontal (Broca's area) and left superior temporal (in this case STS) regions may then be accessed as a function of these projections.[1] The implication of such a route is clearly one driven by top-down processing of the stilled image for the purposes of subsequent internal reconstruction of the frame within a more naturally occurring dynamic format. This is consistent with the subjective experience reported by participants of seeing a stilled speech frame, and internally reconstructing a possible dynamic syllable that would include the observed mouth shape. This route may be implicated in reports of effects of stilled speech images on auditory language processing (Whalen et al., in press), and possibly for the finding that under auditory-imagery-inducing conditions, even written material can access the auditory cortex (Haist et al., 2001).

By contrast, natural time-varying speech is likely to have more direct access to the auditory cortex, via extensive activation of visual movement processing regions (BA 37/19) and the integration of related visual form and movement processing in STS (Puce, Castiello, Syngeniotis, & Abbott, 2001). Dynamic visual-speech displays probably access representations of speech in auditory areas without symbolic mediation of the sort outlined above for stilled images. Previous fMRI evidence suggests that when heard and seen speech are processed as congruent streams, there is preferential activation of the left STS (Calvert et al., 2000). The present study suggests that the correlated time-varying characteristics of vision and hearing in normal audiovisual speech may be one feature of activation in the (left) STS. The resulting perceptual gain achieved when both channels share time-varying characteristics may be realized by subsequent amplification of the signal intensity in the relevant sensory-specific cortices (auditory and visual) via back projections to these areas from the STS (Calvert et al., 1999; Sams et al., 1991). That is, in hearing people, natural silent speech makes use of circuits that have developed for the supramodal processing of speech that is both seen and heard (Calvert et al., 1997; de Sa & Ballard, 1997).

This interpretation is consistent with the dissociated pattern observed in patients with posterior cortical damage and distinctive difficulties in reading seen speech as a function of its movement characteristics. Patient HJA, who had bilateral damage in occipital regions, but spared MT/MST, showed normal sensitivity to natural visible speech, including sensitivity to McGurk effects. He was blind to images of stilled speech. Patient LM, with the opposite pattern of cortical sparing and impairment, was unable to process natural speech (which she found ''disturbing to look at,'' even 17 years poststroke), but she could nevertheless identify stilled images of speech patterns (Campbell, 1992; Campbell et al., 1997).

Both stilled and moving images of speech can afford the perception of phonemic structure, and the cortical circuitry suggests extensive common processing. Image structure, as well as image movement, must contribute to influences of visual on auditory speech and to normal bimodal speech processing. Nevertheless, reliable activation of auditory cortex, and extensive activation of (left) Wernicke's and Broca's region by naturally moving visible speech suggests that it is in its (natural) time-varying characteristics that the core perceptual processes for speech are embodied—both functionally and cortically.

METHODS

Subjects

Eight healthy right-handed subjects (6 men and 2 women, mean age 26, range 22–34), with English as their first language, participated in the study. All subjects were in good health with no history of neurological disorder and gave written informed consent to the protocol that had been approved by the Central Oxford Research Ethics Committee. All subjects had normal or corrected-to-normal (with contact lenses) vision.

Experimental Procedure

Prior to scanning, all subjects were familiarized with the stimuli and task instructions. Subjects were then placed in the scanner, wearing prism glasses that allowed them to view the visual stimuli presented on a projection screen at the end of the scanner bed. Subjects were provided with earplugs and sound-attenuating headphones that reduced the scanner noise to 90 dB. During the scan, subjects were exposed to 30-sec alternating epochs of (A) 10 closed mouth images, (B) 10 stilled images of speech or (C) 10 dynamic consonant + vowel syllables (e.g., arv, boo, eem, sha) in an ABAC block design lasting 10 min. Within each block, all stimuli were presented for 1.5–2.5 sec and interleaved with a skin-toned screen, matched for mean luminosity to avoid flicker-related activations (see Figure 1). In all conditions, the facial image comprised the lower half face (including tip of nose to below the chin) to avoid activations due to eye movements. In the stilled condition, there were 10 frames of each of the stilled images, corresponding to a sustained vowel or a labial or labiodental consonant. The sequence of interleaves and stilled image was piloted to ensure that no apparent movement between image sequences was observed. This frame sequence inhibited apparent movement between adjacent still frames.

The task was phoneme detection. Subjects were instructed to maintain fixation on the mouth area, in order to detect a target ''V'' (e.g., ''voo'' ''ahv'') among other phonemes (“sha,” “eem,” etc.), by button press. In the stilled frame condition the target was stilled at the point of articulating a “v” (visible labiodental gesture—“f-tuck”). The number of target tokens was matched in all conditions. In the closed mouth condition, a written “V” was superimposed on the lip area of the static face image. Button responses were recorded using FMRIB’s Enhanced Stimulation Tool (FEST).

Scanning Protocol

Functional imaging data were acquired on a 3.0T Varian INOVA MRI system with a multislice gradient-echo EPI sequence ($\text{TR} = 2000$ msec; $\text{TE} = 30$ msec, flip angle $= 75°$. $\text{FOV} = 256\ \text{mm}^2$, matrix $= 64^2$) at the Oxford FMRIB Center. Twenty 5-mm-thick axial slices covering the whole brain were acquired every 2 sec over a total scanning period of 10 min. A T1-weighted anatomical scan with a nominal slice thickness of 2.5 mm was also acquired for each subject to aid registration of the individual EPI scans into Talairach space.

Image Analysis

The image analysis was carried out using FMRIB’s Easy Analysis Tool (FEAT) an extension of MEDx (Sensor Systems, VA, USA). Each subject’s EPI data underwent the following prestatistics processing: 3-D motion correction using MCFLIRT (Jenkinson & Smith, 2001); spatial smoothing using a Gaussian kernel of FWHM 5 mm; mean-based intensity normalization of all volumes by the same factor; and nonlinear highpass temporal filtering at a period of 180 sec. Statistical analysis was carried out using FMRIB’s Improved Linear Model (FILM) with local autocorrelation correction (Woolrich, Ripley, Brady, & Smith, 2001). Z (Gaussianized T) statistic images were thresholded using clusters determined by $Z > 2.3$ and a cluster significance threshold of $p = .01$. Registration of the individual EPI images onto an averaged high-resolution T1 brain in Talairach space was carried out using linear registration FLIRT (Jenkinson & Smith, 2001). Fixed-effects group analysis was carried out using FEAT. Z statistic images were first thresholded using clusters determined by $Z > 4.7$ and a cluster significance threshold of $p = .01$.

Acknowledgments

This work was supported by a Fellowship to G. A. C. from the Medical Research Council (MRC) of Great Britain, the FMRIB MRC Core grant, the MRC-IRC at the University of Oxford, and MRC Project Grant G 9702921N. We thank M. J. Brammer for analytic assistance with a pilot study, N. Ramnani and J. Devlin for helpful discussions on the manuscript, and D. Harwood and A. Matheson, Oxford Medical Illustrations, John Radcliffe Hospital, Oxford, for technical support.

Reprint requests should be sent to Gemma A. Calvert, University Laboratory of Physiology, University of Oxford, Parks Road, Oxford OX1 3PT, UK, or via e-mail: gac@physiol.ox.ac.uk.

The data reported in this experiment have been deposited in The fMRI Data Center (http://www.fmridc.org). The accession number is 2-2002-11339.

Note

1. Whether access to MT/MST precedes or follows parieto-frontal engagement is a question for further research.

REFERENCES

Allison, T., Puce, A., & McCarthy, G. (2000). The neurobiology of social cognition. *Trends in Cognitive Sciences, 4,* 267–279.

Bernstein, L. E., Auer, E. T., Jr., Moore, J. K., Ponton, C. W., Don, M., & Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *NeuroReport, 13,* 311–315.

Binkofski, F., Amunts, K., Stephan, K. M., Posse, S., Schormann, T., Freund, H. J., Zilles, K., & Seitz, R. J. (2000). Broca’s region subserves imagery of motion: A combined cytoarchitectonic and fMRI study. *Human Brain Mapping, 11,* 273–285.

Bonda, E., Petrides, M., Ostry, D., & Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neuroscience, 16,* 3737–3744.

Brunet, E., Sarfati, Y., Hardy-Baylé, M.-C., & Decety, J. (2000). A PET investigation of the attributions of intention with a nonverbal task. *Neuroimage, 11,* 157–166.

Buccino, G., Binkowski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R. J., Zilles, K., Rizzolatti, G., & Freund, H.-J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. *European Journal of Neuroscience, 13,* 400–404.

Burnham, D. K. (1993). Visual recognition of mother by young infants: Facilitation by speech. *Perception, 22,* 1133–1153.

Callan, D. E., Callan, A., Kroos, E., & Vatikiotis-Bateson, E. (2000). Multimodal contributions to speech perception revealed by independent component analysis: A single sweep EEG study. *Cognitive Brain Research, 10,* 349–353.

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 cross-modal binding. *NeuroReport, 10,* 2619–2623.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science, 276,* 593–596.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649–657.

Campbell, R. (1992). The neuropsychology of lipreading. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 335,* 39–45.

Campbell, R., Brooks, B., de Haan, E., & Roberts, T. (1996). Dissociating face processing skills: Decisions about lip-read speech, expression and identity. *Quarterly Journal of Experimental Psychology, Section A, 49,* 295–314.

Campbell, R., De Gelder, B., & de Haan, E. H. F. (1996). The laterality of lipreading: A second look. *Neuropsychologia, 34,* 1235–1240.

Campell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology, 32,* 85–99.

Campbell, R., Landis, T., & Regard, M. (1986). Face recognition and lipreading. A neurological dissociation. *Brain, 109,* 509–521.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G. A., McGuire, P. K., Brammer, M. J., David, A. S., & Suckling, J. (2001). Cortical substrates for the perception of face actions: An fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research, 12,* 233–243.

Campbell, R., Zihl, J., Massaro, D. W., Munhall, K., & Cohen, M. M. (1997). Speechreading in the akinotopsic patient, L. M. *Brain, 120,* 1793–1803.

Cathiard, M.-A., & Tiberghien, G. (1994). Le visage de la parole: Une coherence bimodale temporalle ou configurationelle? *Psychologie Francaise, 39,* 357–374.

Cathiard, M.-A., Tiberghien, G., & Abry, C. (1992). Face and profile identification skills for liprounding in normal-hearing French subjects. *Bulletin de la Communication Parlée, 2,* 43–58.

Crozier, S., Sirigu, A., Lehericy, S., van der Moortele, P.-F., Pillon, B., Grafman, J., Agid, Y., Dubois, B., & LeBihan, D. (1999). Distinct prefrontal activations in processing sequence at the sentence and script level: An fMRI study. *Neuropsychologia, 37,* 1469–1476.

Decety, J., & Grezes, J. (1999). Neural mechanisms subserving the perception of human actions. *Trends in Cognitive Sciences, 3,* 172–178.

de Sa, V. R., & Ballard, D. H. (1997). Perceptual learning from cross-modal feedback. *Psychology of Learning and Motivation, 36,* 309–351.

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in-and-out-of-synchrony. *Cognitive Psychology, 11,* 478–484.

Dodd, B., & Burnham, D. (1988). Processing speechread information. New reflections on speechreading. *Volta Review, 90,* 45–60.

Fowler, C. A., & Rosenblum, L. D. (1991). The perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 33–59). Hillsdale, NJ: Erlbaum.

Friston, K., Worsley, K., Frackowiak, R., Mazziotta, J., & Evans, A. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping, 1,* 214–220.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119,* 593–609.

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

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: The psychology of speechreading and audiovisual speech.* Hove, UK: Psychology Press.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, K. N. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception and Psychophysics, 50,* 524–536.

Grèzes, J., & Decety, J. (2001). Functional anatomy of execution, mental simulation, observation, and verb generation of actions: A meta-analysis. *Human Brain Mapping, 12,* 1–19.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., & Blake, R. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience, 12,* 711–720.

Haist, F., So, G., Wild, K., Faber, T. L., Popp, C. A., & Morris, R. D. (2001). Linking sight and sound: fMRI evidence of primary auditory cortex activation during visual word recognition. *Brain and Language, 76,* 340–350.

Hasnain, M. K., Fox, P. T., & Woldorff, M. G. (1998). Intersubject variability of functional areas in the human visual cortex. *Human Brain Mapping, 6,* 301–315.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4,* 223–233.

Hermsdörfer, J., Goldenburg, G., Wachsmuth, C., Conrad, B., Ceballos-Baumann, A. O., Partenstein, P., Schwaiger, M., & Boecker, H. (2001). Cortical correlates of gesture processing: Clues to the cerebral mechanisms underlying apraxia during the imitation of meaningless gestures. *Neuroimage, 14,* 149–161.

Howard, R., Brammer, M. J., Wright, I., Woodruff, P. W. R., Bullmore, E. T., & Zeki, S. (1996). A direct demonstration of functional specialisation within motion-related visual and auditory cortex of the human brain. *Current Biology, 6,* 1015–1019.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science, 286,* 2526–2528.

Jeffers, J., & Barley, M. (1971). *Speechreading (lipreading)*. Springfield, IL: Thomas.

Jenkinson, M., & Smith, S. M. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis, 5,* 143–156.

Jordan, T. R., & Bevan, K. M. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audio-visual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 388–403.

Jordan, T. R., McCotter, M. V., & Thomas, S. M. (2000). Visual and audio-visual speech perception with color and gray scale facial images. *Perception and Psychophysics, 62,* 1394–1404.

Jordan, T. R., & Sergeant, P. C. (2000). Effects of distance on visual and audio-visual speech recognition. *Language and Speech, 43,* 107–124.

Kanwisher, N., McDermott, J., & Chun, M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for the perception of faces. *Journal of Neuroscience, 17,* 4302–4311.

Kanwisher, N., & Moscovitch, M. (2000). The cognitive neuroscience of face processing: An introduction. *Cognitive Neuropsychology, 1/2/3,* 1–13.

Kourtzi, Z., & Kanwisher, N. (2000). Activation in human MT/MST by static images with implied motion. *Journal of Cognitive Neuroscience, 12,* 48–55.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218,* 1138–1141.

Levänen, S. (1999). Neuromagnetic studies of human auditory cortex function and reorganization. *Scandinavian Audiology, 27,* 2–6.

Lewkowicz, D. J. (1996). Infants' response to the audible and visible properties of the human face: I. Role of lexical-syntactic content, temporal synchrony, gender, and manner of speech. *Developmental Psychobiology, 32,* 347–366.

Lewkowicz, D. J. (1998). Infants' response to the audible and visible properties of the human face: II. Discrimination of differences between singing and adult-directed speech. *Developmental Psychobiology, 32,* 261–274.

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences, 4,* 187–196.

Ludman, C. N., Summerfield, A. Q., Hall, D., Elliott, M., Foster, J., Hykin, J. L., Bowtell, R., & Morris, P. G. (2000). Lip-reading ability and patterns of cortical activation studied using fMRI. *British Journal of Audiology, 34,* 225–230.

MacSweeney, M., Amaro, E., Calvert, G. A., Campbell, R., David, A. S., McGuire, P. K., Williams, S. C. R., Woll, B., & Brammer, M. J. (2000). Activation of auditory cortex by silent speechreading in the absence of scanner noise: An event-related fMRI study. *NeuroReport, 11,* 1729–1734.

MacSweeney, M., Campbell, R., Calvert, A., McGuire, P. K., David, A. S., Suckling, J., Andrew, C., Woll, B., & Brammer, M. J. (2001). Dispersed activation in left temporal cortex for speechreading in congenitally deaf people. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 268,* 451–457.

MacSweeney, M., Woll, B., Campbell, R., McGuire, P. K., David, A. S., Williams, S. C. R., Suckling, J., Calvert, G. A., & Brammer, M. J. (2002). Neural systems underlying British Sign Language and audiovisual English processing in native users. *Brain, 125,* 1583–1593.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge: MIT Press.

Massaro, D. (1999). Speechreading: Illusion or window into pattern recognition? *Trends in Cognitive Sciences, 3,* 310–317.

Massaro, D. W., & Cohen, M. M. (1996). Perceiving speech from inverted faces. *Perception and Psychophysics, 58,* 1047–1065.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Munhall, K. G., & Vatikiotis-Bateson, E. (1998). The moving face during speech communication. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 123–139). Hove, UK: Psychology Press.

Olson, I. R., Gatenby, J. G., & Gore, J. C. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Brain Research. Cognitive Brain Research, 14,* 129–138.

Passingham, R. (1993). *The frontal lobes and voluntary action.* Oxford, UK: Oxford University Press.

Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience, 18,* 2188–2199.

Puce, A., Castiello, U., Syngeniotis, A., & Abbott, D. (2001). The human STS integrates form and motion. *Neuroimage, 13,* S931.

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

Rizzolatti, G., & Arbib, M. A. (1999). From grasping to speech: Imitation might provide a missing link. *Trends in Neurosciences, 22,* 152.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Research. Cognitive Brain Research, 3,* 131–141.

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech and Hearing Research, 39,* 1159–1170.

Rosenblum, L. D., & Saldaña, H. M. (1996). An audio-visual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 318–331.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception and Psychophysics, 59,* 347–357.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S.-T., & Simola, J. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters, 127,* 141–145.

Sams, M., & Levänen, S. (1996). When and where are the heard and seen speech integrated? In D. G. Stork & M. E. Henneke (Eds.), *Speechreading in humans and machines* (pp. 232–238). Berlin: Springer-Verlag.

Schweinberger, S. R., & Soukup, G. R. (1998). Asymmetric relationships among perceptions of facial identity, emotion, and facial speech. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1748–1765.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception and Psychophysics, 59,* 73–80.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics, 21,* 427–444.

Senior, C., Barnes, J., Giampetro, V., Simmons, A., Bullmore, E. T., Brammer, M., & David, A. S. (2000). Functional neuroanatomy of implicit motion perception or representational momentum. *Current Biology, 10,* 16–22.

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

Summerfield, A. Q. (1987). Some preliminaries to a comprehensive account of audiovisual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lipreading* (pp. 3–51). Hillsdale, NJ: Erlbaum.

Summerfield, A. Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 335,* 71–78.

Surguladze, S., Calvert, G., Brammer, M. J., Campbell, R., Bullmore, E. T., & David, A. S. (2001). Audio-visual speech perception in schizophrenia: An fMRI study. *Psychiatry Research: Neuroimaging, 106,* 1–14.

Werker, J. F., Frost, P., & McGurk, H. (1992). Cross-language influences on bimodal speech perception. *Canadian Journal of Psychology, 46,* 551–568.

Whalen, D. H., Irwin, J. R., & Fowler, C. A. (in press). A sex difference in audio-visual integration of speech.

Woolrich, M., Ripley, B., Brady, J., & Smith, S. (2001). Temporal autocorrelation in univariate linear modelling of fMRI data. *Neuroimage, 14,* 1370–1386.

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