PHILOSOPHICAL
TRANSACTIONS
OF
THE ROYAL
SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 1001–1010
doi:10.1098/rstb.2007.2155
Published online 7 September 2007

The processing of audio-visual speech: empirical and neural bases

Ruth Campbell*

Department of Human Communication Science, University College London, Chandler House,
2 Wakefield Street, London WC1N 1PF, UK

In this selective review, I outline a number of ways in which seeing the talker affects auditory perception of speech, including, but not confined to, the McGurk effect. To date, studies suggest that all linguistic levels are susceptible to visual influence, and that two main modes of processing can be described: a *complementary* mode, whereby vision provides information more efficiently than hearing for some under-specified parts of the speech stream, and a *correlated* mode, whereby vision partially duplicates information about dynamic articulatory patterning.

Cortical correlates of seen speech suggest that at the neurological as well as the perceptual level, auditory processing of speech is affected by vision, so that 'auditory speech regions' are activated by seen speech. The processing of natural speech, whether it is heard, seen or heard and seen, activates the perisylvian language regions (left > right). It is highly probable that activation occurs in a specific order. First, superior temporal, then inferior parietal and finally inferior frontal regions (left > right) are activated. There is some differentiation of the visual input stream to the core perisylvian language system, suggesting that complementary seen speech information makes special use of the visual ventral processing stream, while for correlated visual speech, the dorsal processing stream, which is sensitive to visual movement, may be relatively more involved.

Keywords: speech reading; audiovisual speech processing; visual speech

1. INTRODUCTION

> Language most shows a man: speak, that I might see thee!
> (Ben Jonson 1572–1637; Timber: Or Discoveries, 1640)

That speech has visible as well as auditory consequences, and that watching the talker can be beneficial for speech understanding, has been acknowledged for many years. Fifty years ago, it was established that seeing the talker could give an improvement in the comprehension of auditory speech in noise equivalent to that produced by an increase of up to 15 dB in signal-to-noise ratio (Sumby & Pollack 1954). This was widely interpreted to mean that the effects of vision on audition were only apparent at low signal-to-noise (S–N) ratios. However, a re-examination of Sumby and Pollack's findings (Remez 2005) clearly shows that the benefit of seeing the talker was not limited to adverse acoustic conditions, but was apparent at all S–N ratios. Nevertheless, throughout the 1960s and 1970s, the impression was that in order for vision to affect speech perception, acoustic information needed to be suboptimal.

Things changed in the 1970s and 1980s. First came the demonstration that the perception of certain speech segments could be strongly influenced by vision even when acoustic conditions were good, and indeed that some audio-visual pairings could lead to illusory perceptions. The original discovery of the McGurk effect (McGurk & MacDonald 1976) was accidental. The investigators were researching young children's imitation of auditory speech patterns. They dubbed a number of different video to auditory syllable tokens with the aim of distracting the child from the auditory imitation task. The syllables had varied consonant–vowel forms, including ba, ga, da, ka, ta and pa. When a seen 'ga' was dubbed to a heard 'ba', all participants thought that 'da' had been said—and the technician was reprimanded for not dubbing the 'ga' and 'ba' correctly. Only when he insisted that the tokens had the required form, and participants tested their perceptions by closing their eyes and watching the silent videotape closely, did it become apparent that 'da' was illusory. That is, under these specific conditions, the perceiver heard an event which was not present in either the visual or the auditory stimulus. The illusion also held for the unvoiced synthesis (visual 'ka'; auditory 'pa'… hear 'ta'). It was as marked for children as for adults and was found to be relatively insensitive to knowledge of its bases or to lexical or other expectations. The McGurk illusion thus added a new impetus to studies of audio-visual speech.[1] In another set of studies, Reisberg *et al.* (1987), using natural rather than dubbed audio-visual speech, and extended passages of speech rather than isolated tokens, reported that, even when hearing conditions are excellent, there is a gain in speech comprehension under audio-visual compared with auditory-alone conditions. This occurred for hard-to-understand but easy-to-hear passages. What is it about audio-visual processing that can deliver such outcomes? How does vision affect the primary modality of audition for understanding speech?

*r.campbell@ucl.ac.uk

One contribution of 13 to a Theme Issue 'The perception of speech: from sound to meaning'.

This paper presents some experimental findings, including behavioural and neurological data, that explore the idea that understanding speech requires that we take into account its visual concomitants. I will consider both speech-reading in the absence of hearing (silent speech-reading) and audio-visual speech perception, and will offer some suggestions concerning multimodal mechanisms.

2. THE SOURCE–FILTER MODEL OF SPEECH: SOME APPLICATIONS TO SPEECH-READING

A model of speech production can be constructed based on the physical characteristics of the system that is used to produce speech (Fant 1960; Diehl 2008). One source of the initial acoustic event is the vibration of the vocal folds (the rate of vibration determining fundamental voice pitch); the filter function describes the effects on the resultant acoustic waveform of its passage through the rest of the vocal apparatus—the vocal cavities, the hard and soft palates, the tongue and mouth. It may be possible to detect some visual correlates of source function. For example, Munhall *et al.* (2004) report that in sentential utterances, head movements are quite well temporally aligned with the onset and offset of voicing—and hence there are correspondences between the kinematics of head actions and the dynamic sound pattern over the period of the utterance as seen in the speech spectrum. When we speak, our vocal folds do not function independently of other bodily actions, as you can confirm for yourself when watching a talker from behind. The onset and offset of speech, in particular, are relatively easy to detect from head movements.

As well as source effects, many aspects of the filter function are visible. In women and children, the length of the vocal tract is generally shorter than in men. Gender and age are predominantly identified by sight. As for the configuration of the vocal tract, mouth opening and closure, as well as mouth shape, are all highly visible. Visible configurations of the lips, teeth and tongue allow us to distinguish ‘map’ from ‘nap’, ‘threat’ from ‘fret’, ‘tap’ from ‘tack’ and ‘him’ from ‘ham’ by eye. While place of articulation can often be determined visually, manner of articulation can also sometimes be seen: for instance, the late voicing of ‘p’ in ‘park’ can be accompanied by a visible lip-puff, which is absent when ‘bark’ is uttered.

Source–filter models are appropriate for the description of speech production, and can account for some specific visual as well as acoustic properties of speech. But are these just a few local features, or do speech production characteristics have broader applicability to speech-reading? Yehia *et al.* (2002) measured the visual kinematics as well as the spectral (acoustic) properties of some spoken phrases. The kinematics of the talker’s face and head were correlated with spectral events in these utterances, to the extent that the visible motion characteristics could be used to estimate and predict (i.e. recover) almost all of the speech acoustic patterns. Movements of the mouth and lips, while contributing to the synthesis, were not themselves as useful as head, face (eyebrows especially) and mouth movements. Such demonstrations suggest that purely visual spatio-temporal speech patterns afford reliable access to representations of phrase-length utterances, and, moreover, that speech-reading may usefully consider actions of the face and head beyond those of the mouth and lips.

Summerfield (1987) suggested that when we perceive speech, we reconstruct the patterns of articulation used by the talker, irrespective of the modality of input. Visual and even haptic processes (Fowler & Dekle 1991) can affect the impression of what was heard. While information from these modalities may be integrated with acoustic information via purely associative mechanisms, it seems probable that the processing system will make use of the correspondences between visual, somaesthetic and acoustic events to inform processing. Since these are all the consequences of the act of speaking, implicit knowledge about articulatory processes is likely to influence multimodal speech processing. Naturally, if we hear well, speech representations will be dominated by our acoustic impressions, but, nevertheless, speech perception cannot be considered to be exclusively auditory. Many explorations of how non-acoustic impressions of the talker can moderate segmental speech perception have been undertaken. Green and colleagues performed some of the most convincing of these. Among other things, Green *et al.* (1991) showed that the visual impression of a talker’s gender could shift the perception of a clearly heard but ambiguous auditory consonant from ‘sh’ to ‘s’. ‘s’ is produced with the tongue immediately behind the teeth, while for ‘sh’ the place of articulation is more posterior. The perceiver can gain an impression of the vocal-tract characteristics of the talker by vision alone; in this case, the estimation is of the place of articulation in relation to the probable depth of the mouth cavity.

Whether speech is considered at the ‘fine-grain’ level of phonetic context for phoneme discrimination, or the ‘coarse-grain’ level of the spectral characterization of a 2 or 3 s utterance of connected speech, there is good evidence that the talker’s seen actions can contribute to the perception of speech.

3. BINDING: SOME PRELIMINARIES

It is one thing to claim that articulatory events can be perceived amodally (or supramodally, or cross-modally), but quite another to describe exactly how, when one perceives natural speech from a talker who one sees as well as hears, visual and auditory information may combine to allow the speech processor to select the appropriate fit. What is required for an audio-visual speech event to be processed? First, is attention needed or is audio-visual processing automatic and mandatory? It has long been claimed that McGurk effects are automatic. McGurk effects do not require attention to be explicitly directed to them to be experienced (e.g. Soto-Faraco *et al.* 2004). Infants who are not yet able to speak or respond to attentional instruction are sensitive to audio-visual fusions. From the age of six months or so, infants who have habituated to a ‘McGurk’ audio-visual ‘da’ (the stimulus comprises an auditory ‘ba’ dubbed to a visual ‘ga’) dishabituate when a congruent audio-visual ‘ba’ is played to them, and fail to respond with a dishabituation response when a 'real' audio-visual 'da', derived from a visual 'da' and an auditory 'da', is played to them (Burnham & Dodd (2004) and see also Rosenblum *et al.* (1997)). Audio-visual speech processing capabilities are in place even before the child has useful speech.[2] Nevertheless, studies with adults suggest that there can be an attentional cost to processing audio-visual fusions of this sort. Tiippana *et al.* (2004) and Alsius *et al.* (2005) independently showed that a visual distractor task could reduce the incidence of illusory McGurk percepts. In Tiippana *et al.*'s study, the distractor element was a moving image of a leaf randomly floating across the face of the talker (but not obscuring the mouth). In the Alsius *et al.* experiment, line drawings of objects were overlaid on the image of the talker. In both cases, it might be claimed that the visual distractor degraded the visual input, so the reduction of the McGurk effect resulted from perceptual rather than attentional processes. However, Alsius *et al.* (2005) showed that *auditory* distraction during the presentation of McGurk stimuli also reduced vulnerability to McGurk effects. Thus, paradoxically, adding an auditory task to the identification of an audio-visual syllable *increased* the probability of an auditory response to an incongruent audio-visual item. This strongly suggests that it is the process of integrating vision and hearing which is vulnerable to additional attentional load, rather than the processing of each input stream prior to their integration.

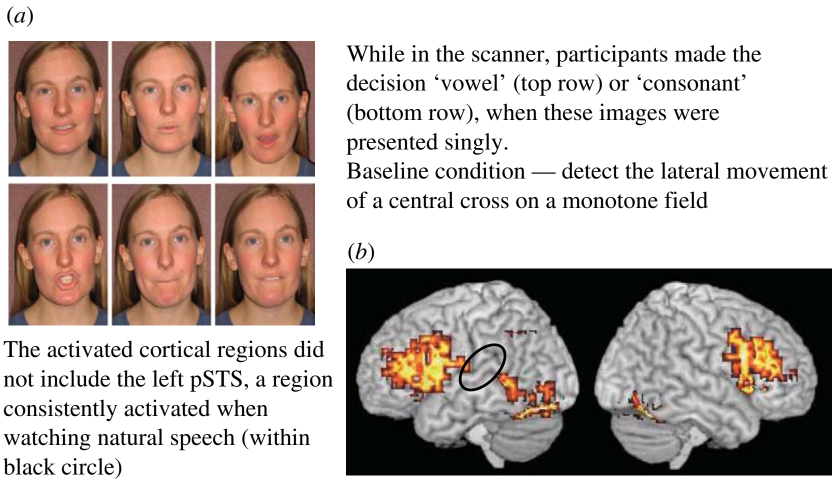
In McGurk-type experiments, participants are presented with well-synchronized visual and auditory tokens, apparently emanating from a single talker; that is, the auditory and visual parts of the McGurk stimuli are spatially and temporally coherent and coextensive. It might be thought that this is a strong cue to their co-processing, to their 'binding'. In fact, audio-visual effects, including McGurk effects and an audio-visual advantage for bimodal compared with purely auditory processing, are reported under relatively large desynchronizations. The effects of vision on audition occur for asynchronies (vision leading) of 250 ms or more, depending on the task (Grant *et al.* 2004). To some extent, the loose temporal fit of vision to aftercoming sound may reflect anticipatory coarticulation—appropriate movements of the mouth often occur prior to vocalization. Certainly, tolerance of vision-led asynchronies is greater than for audition-led asynchronies.

In addition to tolerance of asynchronies, displacement of a voice in space is not accurately perceived. Wherever its actual source, perceivers locate an artificial speech source at the position of a visually perceived apparent talker (the ventriloquism illusion), i.e. vision 'captures' the location of the auditory event (Radeau & Bertelson 1974). When the auditory channel comprises two 'overlaid' voices, heard to be speaking simultaneously, and they issue from a single central loudspeaker, the perceiver not only uses a video display corresponding to one of the utterances to locate and shadow that talker effectively, but, surprisingly, may also be more able to shadow the *unseen* talker, who is now perceived to be spatially separated from apparent location of the visible talker (Driver 1996).

It seems that the speech processor is relatively unconcerned about the fine spectro-temporal and locational details of the match between vision and audition. Any model of the pattern-matching system that binds auditory and visual speech events must take account of this looseness of fit. The looseness may arise from the relative dominance and salience of audition in the perception of speech, so that vision, as the secondary input stream, is not required to match the auditory spectro-temporal markings precisely. A flexible system, using attentional resources where necessary, appears to be at work, allowing any analysis-by-synthesis approach of vision and audition to exercise variable constraints depending on the salience of the input event (cf. Grant *et al.* 2004). An alternative proposal is that, if articulatory plans are incorporated into the representations to which such perceptual events map, they may specify the acoustic and visual correlates of an utterance less with respect to location and temporal patterning and more in relation to the somaesthetic consequences of a speech act.

4. WHAT DOES VISION DELIVER? THE ART OF 'HEARING BY EYE'

The McGurk effect results from perceptual integration of a visible open mouth syllable (e.g. 'da') with a heard one (e.g. 'ga'), which has spectral similarities to the syllable that is perceived following combination (e.g. 'ba'). The combinatorial rules and processes involved have exercised psychologists for many years (see Bernstein *et al.* 2004*a* for review), but this work has focused on processing at the phoneme level. The findings that phonetic context perceived by eye can shift phonemic category boundaries (Green *et al.* 1991), and that prelinguistic babies are sensitive to McGurk effects (Burnham & Dodd 2004), demonstrate that audio-visual integration can occur 'pre-phonemically'. That said, the phonemic level of linguistic structure offers the most approachable entry point for examining many aspects of the perception of seen speech in the *absence* of hearing—that is, silent speech-reading. Some speech-read segments are relatively unambiguous. For instance, labiodental consonants and English point vowels enjoy a high level of audio-visual mapping consistency ('what you see is what you hear'). However, a seen speech event usually maps onto several (acoustically defined) phonological categories. Visually confusable phonemes ('visemes') can be considered to constitute a phonemically equivalent class (PEC; Auer & Bernstein 1997). The number of PECs will vary from person to person, depending on their speech-reading skill and on the visibility of the talker's speech. Auer & Bernstein (1997) found that 12 PECs were sufficient to identify most English words. This number corresponds well with theoretical studies, suggesting that this number of distinctions should suffice for useful visual speech-reading as an aid to hearing, and contrasts with estimates of approximately 40 phonemes available 'by ear' in spoken English. The reason why a relatively small number of PECs can, in principle, suffice for identifying individual spoken words is that most words in English are relatively unique in their segmental and syllabic structure. That is, lexical space in English is relatively sparsely occupied and well distributed. Heard speech can, on this type of analysis, be considered to be overdetermined, containing a great deal of structural redundancy. This perspective allows us to understand the



While in the scanner, participants made the decision 'vowel' (top row) or 'consonant' (bottom row), when these images were presented singly.
Baseline condition — detect the lateral movement of a central cross on a monotone field

The activated cortical regions did not include the left pSTS, a region consistently activated when watching natural speech (within black circle)

campbell audiovisual speech

Figure 1. (*a*) Speech images used in the fMRI experiment reported by Capek *et al.* (2005). Vowels are shown in the top row and consonants in the bottom row. (*b*) Rendered group activation maps for the task of distinguishing vowel and consonant lip shapes. Images were presented singly for decision. The baseline task was to detect the movement of a cross on a blank background (Capek *et al.* 2005). pSTS (black circle) was not activated. Significant foci of activation (x, y, z, coordinates) included: (i) inferior temporal cortex/fusiform gyrus (−29, −78, −17), (ii) right inferior frontal cortex extending into dlpfc (47, 11, 26), (iii) left inferior frontal cortex extending into dlpfc (−47, 7, 33), (iv) left inferior parietal lobule (−25, −63, 43), and (v) caudal anterior cingulate gyrus (−3, 11, 50).

finding (Reisberg *et al.* 1987) that seeing as well as hearing the talker improves understanding of conceptually difficult texts, even under excellent perceptual conditions. We can assume that understanding such texts requires allocation of limited cognitive resources. If the audio-visual speech signal is highly redundant, it may require minimal cognitive processing to determine what words were spoken, thus freeing cognitive resources to allow better interpretation of the utterance.

The redundancy of speech can also explain why some people attain good speech comprehension by sight alone—at least under optimal talking and viewing conditions—as demonstrated by 'super-speech-readers' (e.g. Andersson & Lidestam 2005). These are people, often deaf from an early age, whose speech-reading abilities allow them to follow silently spoken conversations with high accuracy. If just 12 PECs are required to identify more than 90% of English words, one can understand how, in principle, such accuracy can be achieved by speech-reading alone—especially given that higher-level constraints (topic, discourse constraints, syntax and meaning) can also be used to aid comprehension. From this perspective, the fact that most hearing people are relatively poor speech-readers may reflect relative (over-) reliance on acoustic parameters of the speech stream.

5. THE VISIBLE SPEECH STREAM: VARIETIES OF INFORMATION

An interesting feature of speech is that segments that are confusable acoustically (for instance, 'm' and 'n', and 'th' and 'f') are often visually distinctive—and vice versa ('p' and 'b' are acoustically distinct, but visually confusable; see Summerfield (1987) for illustrations). While this may implicate vision in some aspects of the evolution of spoken languages, a more pressing concern is to try to answer the question—what are the visual features of such distinctive speech segments? Summerfield (1979) showed that speech-reading relied mainly on mouth shape, mouth opening and the visible position of the tongue (and sometimes teeth). Is this just because we are disposed to perceive a simple correspondence between the diameter of a probable sound source and its amplitude? This cannot be a critical feature, as the identification of speech in noise is not helped by the perception of an annulus whose diameter is controlled by the amplitude of the acoustic signal (also see Bernstein *et al.* 2004*b*; Ghazanfar *et al.* 2005). This (among other demonstrations) suggests that, for the most efficient speech-reading, mouth opening and closing and the tongue position should be clearly visible. Figure 1 indicates how still images of faces producing speech can be reliably classified in terms of their speech characteristics. Vowels and consonants can be easily identified from a closed set, when mouth shape and the configuration of lips, teeth and tongue are visible.

A number of demonstrations, however, suggest that the configuration of the mouth, tongue and teeth, as captured in a still image, may not fully explain the audio-visual advantage or account fully for McGurk effects. McGurk effects can be obtained at viewing distances too great for mouth disposition to be clearly visible (Jordan & Sergeant 2000). Rosenblum and his colleagues have used point-light-illuminated faces to explore the time-varying aspects of seeing speech. Typically, 12–20 points on the face surface, videotaped at normal speed, give information about dynamic deformation of the face surface in the absence of any facial features. Such sparse stimuli induce McGurk effects (albeit at a reduced level compared with full facial images; Rosenblum & Saldaña 1996) and an audio-visual gain for speech perceived in noise (Rosenblum *et al.* 1996). Another manipulation that differentially affects the visibility of specific face features is spatial frequency filtering of the image. As long as the temporal characteristics of the signal are maintained, low-pass spatial frequency filtering

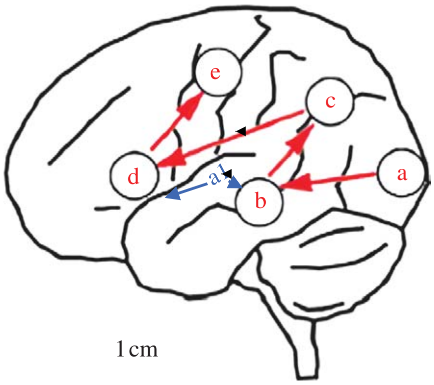


Figure 2. Schematic of the left hemisphere showing locations and activation sequence for the processing of visual speech (adapted from Nishitani & Hari (2002)). Participants in this MEG study of silent speech-reading identified vowel forms from videoclips. The following regions were activated, in sequence (a) visual cortex, including visual movement regions, (b) superior temporal gyrus (secondary auditory cortex), (c) pSTS and inferior parietal lobule, (d) inferior frontal and (e) premotor cortex. Auditory inputs (primary auditory cortex, A1) are hypothesized to access this system at (b).

(down to 19 cycles per face) blurs the image without markedly compromising audio-visual gain for understanding spoken sentences (Munhall *et al.* 2004). Visible kinematics can deliver critical components of the audio-visual advantage. Also critical to useful visual speech is the temporal sampling rate. When this is below 10 frames s^{-1}, the audio-visual advantage may decrease or disappear. The average rate of opening and closing of the vocal tract in normal speech is approximately 12 Hz. This is likely to be the minimal sampling rate for effective audio-visual speech processing.

6. COMPLEMENTARITY AND REDUNDANCY IN THE SPEECH STREAM

To summarize, audio-visual processing is more effective than auditory processing of natural speech for two reasons. First, some segmental contrasts can be seen clearly, thus aiding speech comprehension, especially where those segments are acoustically confusable. Second, many features of an utterance can be perceived by both ear and eye: the audible and the visible patterns are highly correlated, reflecting the underlying dynamics of speech production. The speech processing system makes use of the redundancies offered by the similar pattern of time-varying signal change across the modalities. Thus, there are, in principle, two modes whereby seen speech can affect what is heard: a *complementary* mode, whereby vision provides information about some aspects of the speech event that are hard to hear, and which may depend on the shape and contour of the lower face being clearly visible; and a *correlated* mode, where the crucial feature of the speech stream is its temporo-spectral signature, which will show regions of similar dynamic patterning across both audible and visible channels. This latter mode must require the perception of visible motion, and studies of patients with acquired lesions differentially affecting visual processing of form and motion bear this out (Campbell *et al.* 1997). Different cross-modal binding principles may apply depending on the relative importance of complementary and correlated information in the utterance to be processed.

7. NEURAL MECHANISMS FOR AUDIO-VISUAL AND VISUAL SPEECH

Reviews of the neural bases of visual and audio-visual speech processing can be found elsewhere (e.g. Callan *et al.* 2004; Calvert & Lewis 2004; Capek *et al.* 2004; Miller & D’Esposito 2005). Some well-established findings are itemized here, which will then be examined in relation to the ‘two-mode’ sketch outlined in §5. They apply to hearing people: the special case of speech-reading in profound prelingual deafness is different again (Sadato *et al.* 2002).

Before summarizing these findings, it is worth pointing out that many aspects of the patterns of brain activation to be described here may also be found in non-human primates, when animals are presented with species-specific calls unimodally or bimodally. Since distinctive activation patterns can be traced in the brains of such non-speaking species—and regions homologous to those in humans appear to be implicated—within-species communication, rather than speech-specific, mechanisms may underlie many of the findings related to multimodal audio-visual speech processing (Ghazanfar *et al.* 2005).

(i) Speech-reading in the absence of any auditory input (silent speech-reading) activates auditory cortex. This may include activation within core regions of primary auditory cortex (A1; Pekkola *et al.* 2005), although the extent and specificity of activation within auditory cortex is problematic. While all investigators agree that parts of the superior temporal plane adjoining the upper part of the superior temporal gyrus are activated by silent speech, there had been disagreement concerning the extent to which primary auditory cortex within Heschl’s gyrus might be activated by seen silent speech (Calvert *et al.* 1997; Bernstein *et al.* 2002). Pekkola *et al.*’s findings, using more powerful scanning techniques than earlier studies, have unambiguously demonstrated that primary auditory cortex can be activated by silent speech-reading. Now what needs to be determined is the specificity of activation in A1 to speech-like events.

(ii) Speech-reading tends to generate left-lateralized or bilateral activation (e.g. Calvert & Lewis 2004; Capek *et al.* 2004). This is in contrast to the usual finding for other face actions, such as perception of gaze direction or facial expression, which tend to show more extensive right-lateralized activation.

(iii) The middle and posterior parts of the superior temporal gyrus, including the posterior superior temporal sulcus (pSTS), are reliably and consistently activated by silent speech-reading, and also by audio-visual speech. This region usually constitutes the principal focus of activation in fMRI studies of speech-reading (e.g. Calvert *et al.* 1997, 2000; Ludman *et al.* 2000; MacSweeney *et al.* 2002; Wright *et al.* 2003; Callan *et al.* 2004;

Capek *et al.* 2004; Hall *et al.* 2005; Skipper *et al.* 2005).

(iv) (Left) pSTS can (but does not always) show differential activation for congruent and incongruent audio-visual speech. It can show supra-additive activation for (congruent) audio-visual speech compared with unimodal seen or heard speech (e.g. Calvert *et al.* 2000; Wright *et al.* 2003; Miller & D'Esposito 2005). Inhibitory (i.e. sub-additive) activation for audio-visual compared with unimodal input can be observed in other parts of the superior temporal gyrus (e.g. Wright *et al.* 2003), and for incongruent audio-visual pairings within pSTS. While the findings are variable, they are generally consistent with the idea that pSTS is a primary binding site for audio-visual speech processing.

(v) Inferior frontal regions, including Broca's region (BA 44/45), and extending into anterior parts of the insula, are activated by speech-reading. Often, watching speech generates greater activation in this region than observing other actions or listening to speech (e.g. Buccino *et al.* 2001; Campbell *et al.* 2001; Santi *et al.* 2003; Watkins *et al.* 2003; Ojanen *et al.* 2005; Skipper *et al.* 2005).

8. TIME COURSE OF CORTICAL ACTIVATION

The time course of functional cortical activation has been studied primarily using scalp-recorded event-related potentials (ERP) and magnetoencephalography (MEG), both of which offer online methods of tracking brain events and have good temporal resolution. The results obtained to date suggest the following sequence of events, many of which are illustrated in figure 2.

(i) Following unimodal processing within the relevant primary sensory cortices, superior temporal regions are implicated in binding the different sensory streams in speech events (Miller & D'Esposito 2005). So, for example, a distinctive electrophysiological signature for a McGurk audio-visual stimulus compared with an audio-visual congruent one is likely to reflect a signal source in superior temporal regions—within approximately 150 ms of auditory signal onset (e.g. Sams *et al.* 1991; Colin *et al.* 2002; Möttönen *et al.* 2004). One study using McGurk-type stimuli suggests that the amplitude and latency of the auditory evoked potential related to auditory identification (the N1/P2 complex) can be reduced when vision leads audition slightly (van Wassenhove *et al.* 2005). The better the visual event predicts the following auditory one, the smaller the auditory EP. This waveform and its sensitivity to information in the visual stream probably arise within pSTS.

(ii) Activation in pSTS extends posteriorly to the junction with the parietal lobe. Activation for silent speech-reading, like that for audio-visual speech, then extends anteriorly to inferior frontal regions (Nishitani & Hari 2002), via temporo-parieto-frontal junction activation (figure 2).

(iii) Activation can also project from (secondary) multisensory sites, such as pSTS, to (primary) unimodal visual and auditory processing regions (back projection; Calvert *et al.* 2000). Back projection provides the most likely mechanism for activation in auditory cortex produced by silent speech-reading (see Calvert & Lewis 2004).

(iv) Activation in somatosensory cortex has been reported for silent speech-reading (Möttönen *et al.* 2005). This recent finding adds weight to the consideration of speech perception in terms of all of its multimodal properties.

I have suggested that speech-reading may be considered in terms of two different processing modes: one largely dependent on perceiving the configuration of the mouth, lips and tongue, which can provide complementary visual information to that in the auditory speech stream, and the other reflecting correlations between the kinematics of heard and seen speech and carrying useful informational redundancy in the multimodal speech stream. If this conceptualization is valid, then the different modes may have distinctive cortical activation characteristics. Studies exploring the effects of manipulating the display, and the response task, may offer further insight into this.

9. POSTERIOR SUPERIOR TEMPORAL SULCUS: SPEECH ACTIONS, BUT NOT ALL SPEECH IMAGES

There are hints that pSTS is especially sensitive to dynamic aspects of seen speech. This might mean that visible speech information that drives activation in this region is related primarily to the dynamic aspects of the heard speech stream (correlated mode), rather than to the visibility of specific facial configurations (complementary mode). Callan *et al.* (2004) used spatial filtering to vary the amount of facial detail visible in spoken sentences, presented audio-visually in noise (speech babble). When fine spatial detail was accessible (natural and middle-pass filtered video), there was more activation in the middle temporal gyrus (MTG) than when the video was low-pass filtered. Under low-pass filtering, which reduced the visibility of facial detail, pSTS was activated, while under normal and middle-pass filtering both pSTS and MTG were activated.

Calvert & Campbell (2003) compared activation in response to natural visible (silent) speech and a visual display comprising sequences of still photo images captured from the natural speech sequence. Spoken VCV disyllables were seen. The still images were captured at the apex of the gesture—so for 'th', the image clearly showed the tongue between the teeth, and for the vowels, the image captured was that which best showed the vowel's identity in terms of mouth shape. The still series thus comprised just three images: vowel; consonant; and vowel again. However, the video sequence was built up so that the natural onset and offset time signatures of the vowel and consonant were preserved (i.e. multiple frames of vowel, then consonant and then vowel again). The overall duration of the still lip series was identical to that for the normal speech sample, and care was taken to avoid illusory movement effects as the vowel and consonant images changed, by using visual pink noise frames interleaved with those of the speech series. The visual impression was of a still image of a vowel (approx. 0.5 s), followed by a consonant (approx. 0.25 s), and again followed by a vowel. Participants in the scanner were asked to detect a consonantal target ('v') among the disyllables seen. Although pSTS was activated in both natural and still conditions, it was activated more strongly by normal movement than by the still image series. In a complementary finding, Santi *et al.* (2003) found that point-light-illuminated speaking faces—which carried no information about visual form—generated activation in pSTS. Finally, a recent study (Capek *et al.* 2005, in preparation) used stilled photo images of lip actions, each presented for 1 s, for participants in the scanner to classify as vowels or consonants. Under these conditions, no activation of pSTS was detectable at the group level—nor in individual scans (figure 1). Images of lips and their possible actions are not always sufficient to generate activation of this region; pSTS activation requires that either visual motion be available in the stimulus or the task requires access to a dynamic representation (of heard or seen speech). This sole negative finding concerning the involvement of pSTS underlines its crucial role for most speech-reading and audio-visual perception. It strongly suggests that pSTS is especially involved in the analysis of natural speech whose visual movement characteristics correspond with auditory spectro-temporal features. However, the functional role of pSTS is not confined to multimodal or unimodal speech processing. It is activated extensively in the integration of biological visible form and motion (Puce *et al.* 2003). pSTS is activated when imitating or observing imitations of the actions of others (for review, see Buccino *et al.* 2001; Brass & Heyes 2005). pSTS is also activated by the presentation of learned, arbitrary audio-visual pairings (for discussion, see Miller & D'Esposito (2005)). This must inform theorizing concerning its role in visual and audio-visual speech perception.

10. PHOTOGRAPHS OF SPEECH: MORE ROUTES TO VISIBLE SPEECH PROCESSING AND AUDIO-VISUAL INTEGRATION?

pSTS was not activated when participants examined photographs to decide whether a vowel or consonant was being spoken. There must therefore be networks that can support some aspects of seen speech perception that are not directly reliant on pSTS. Where could these be? Stilled lip images generated relatively more activation than moving lips within primary visual areas V1 and V2 (Calvert & Campbell 2003). These primary visual regions project to the ventral visual system, including inferior temporal cortex. Projections radiate from this region to the middle and (anterior) superior temporal cortex. These parts of the temporal lobe support a range of associative processes—they are traditionally regarded as 'secondary association areas' for categorizing and associating inputs from the senses. These inferior and middle temporal regions of the ventral visual system may have been accessed in the case of the motion-blind patient, LM, who could identify speech sounds that were associated with isolated seen speech photographs, but was unable to identify natural visible speech, and showed no effect of vision on audition when presented with naturally moving McGurk stimuli (Campbell *et al.* 1997). LM had bilateral damage to the lateral occipital cortices, including area V5, which project to pSTS. Here is one means by which complementary information concerning the precise place of articulation may be made available to the cognitive system. It should also be noted that another multimodal region, the left inferior parietal lobule (IPL), a region dorsal to pSTS, which is often activated when people are engaged in segmental speech analysis (Scott 2005), was activated by seen speech in the experiment of Capek *et al.* (figure 1). IPL may be accessed not only by projections from pSTS, but also from other projections, possibly from inferior frontal regions.

11. THE BROADER PICTURE

Current theorizing about auditory speech processing suggests two major processing streams emanating from primary auditory cortex to generate activity in the perisylvian regions (superior temporal and inferior frontal) of the left hemisphere (Scott 2005; Patterson & Johnsrude 2008). One stream runs anteriorly along the upper surface of the temporal lobe. This becomes increasingly sensitive to the semantic characteristics of the utterance (a 'what' stream). The other runs dorsally through the superior temporal gyrus to the temporo-parieto–frontal junction and is especially sensitive to the segmental properties of speech. This may constitute part of a 'how' stream, concerned with the specification of the segmental properties of speech in articulatory and acoustic terms. Both 'how' and 'what' streams project to inferior frontal regions including Broca's area, though through different tracts. One function for the posterior left-lateralized network as a whole, including its projections to frontal regions, may be to 'align' the segmental specifications of speech whether it is planned, produced or perceived (Hickok & Poeppel 2004). That is, the frontal component, related to the planned articulation of speech, and the temporo-parietal component, concerned with the acoustic specification of the speech segment, need to interact to develop representations of segmental speech forms. By contrast, the anterior stream, including its frontal projections, may be differentially specialized for analysing meaning in larger linguistic units (Thompson-Schill 2005).

The picture sketched earlier in this review suggested two modes whereby seen speech may influence the processing of heard speech. One was described as a complementary mode, making use of face information to distinguish speech segments by eye, which may be hard to distinguish by ear. The other is a correlated mode, for which information in the visible speech stream that is dynamically similar to that in the auditory stream provides useful redundancy. My contention is that these are reflected not in completely discrete cortical processing systems, but rather in relatively differentiated access to two major streams for the processing of natural language—a 'what' and a 'how' stream. The 'what' stream makes particular use of the inferior occipito-temporal regions and the ventral visual processing stream, which can specify image details effectively. It can therefore serve as a useful route for complementary visual information to be processed. A major projection of this stream is towards association areas in middle and superior temporal cortex. To the extent that seen speech (whether alone or in combination with sound) activates language meanings and associations, it will engage these anterior and middle temporal regions, and possibly bilaterally rather than left-lateralized. This stream could be accessed effectively by still speech images. When the addition of vision to audition generates changes in meaning, especially at the level of the phrasal utterance, functional activation in these cortical regions could be differentially engaged. By and large, the process whereby visual and auditory characteristics are bound together within this route is primarily associative. That is, learned associations of images and sounds of speech are associated with a specific response pattern. Thus, in the study of Capek *et al.* (2005), the requirement to identify images of vowels and distinguish them from consonants may have used a 'general purpose' associative mechanism. The specificity of associative processing in this region is untested. It may be limited to object-based associations (glass shatters, ducks quack). In contrast to this, the 'how' stream for the analysis of auditory speech may be readily accessed by natural visible speech, characterized by dynamic features that correspond with those available acoustically. Processing that requires sequential segmental analysis (e.g. identifying syllables or words individually or in lists) will differentially engage this posterior stream. It is in this stream that the *correlational* structure of seen and heard speech is best reflected. The visual input to these analyses arises primarily in the lateral temporo-occipital regions that track visual movement, which project primarily to pSTS. pSTS has been shown to play a range of roles in intra- and intermodal processing, but the suggestion here is that it may have a crucial role in processing the supramodal dynamic patterns that characterize natural audio-visual speech by abstracting relevant features from both the visual and the auditory stream. One should be cautious, though, in predicting that this is the only network involved in audio-visual speech binding. Among other things, we do not yet know the extent to which the hypothesized posterior audio-visual stream is responsible for cross-modal integration of vision and audition in the perception of speech prosody, or for the interplay of vision and hearing in the perception of spoken discourse (but see Skipper *et al.* 2005).

When we learn to speak, our developing vocalizations tune to those that we hear—we imitate the vocal patterns of our language teachers. Clearly, hearing other talkers—and matching our own utterances to those we hear—is a crucial part of the development of speech. Yet how we do this, and the relative contributions of the perisylvian regions to the development of amodal representations that capture articulatory as well as auditory and other sensory features, remains mysterious. There is little doubt that our experiences of performing actions leave strong traces in the representations we use to perceive those actions. Following the classical studies of Meltzoff and colleagues, which showed that human neonates imitate visually observed mouth movements (Meltzoff & Moore 1983), the ability to imitate the visually observed actions of others has become the focus of studies that suggest that frontal cortical regions may contain mirror-neuron assemblies, specialized to respond to the perception of particular actions, as well as their planning and execution (Rizzolatti & Arbib 1998). Several studies now confirm that Broca's area within the inferior lateral left frontal lobe, classically understood to be involved primarily in the selection of speech acts for production, is especially active in processing seen speech (e.g. Buccino *et al.* 2001; Campbell *et al.* 2001; Watkins *et al.* 2003; Skipper *et al.* 2005), even when no overt speech action is required. The studies do not, however, 'prove' the mirror-neuron hypothesis, which places the primary perception–action link in a specific frontal region, and whose homology to Broca's area is uncertain. Rather, it seems that input from the primary visual sensory regions drives activation in specific temporal regions that are in turn connected to inferior parietal–inferior frontal circuits. Audio-visual and visual speech perceptions thus bring about cortical activation of action plans and sequences, as well as some somaesthetic consequences of speaking. Does the extent of inferior frontal activation then determine or constrain the *perception* of an audio-visual or a visual speech gesture? There are claims that this is so (Skipper *et al.* 2005; van Wassenhove *et al.* 2005), and there is no doubt that activation in Broca's area can be shown to play a distinctive role in speech-reading and audio-visual speech (Sams *et al.* 2005; Pekkola *et al.* 2006). However, whether such activation is a necessary component of speech perception is unproven. That said, to have non-auditory sense—by sight, by 'feel' and by articulatory knowledge—of both talk and the talker is a vital (although possibly not a sufficient) component of speech mastery.

I am grateful to Cheryl Capek for reading drafts of this paper, and to the Wellcome Trust, the Medical Research Council of Great Britain, the Royal Society of London and the Economic and Social Research Council for financial support related to the work reported here.

ENDNOTES

[1] The McGurk effect is not illusory in the sense that it is a distortion of 'normal perception'. Its illusory nature lies in the specific and unique combination of visual (velar stop consonant ('k','g')) and auditory (bilabial stop consonant ('p','b')) inputs giving rise to an apparent alveolar stop consonant ('t','d'), which did not occur in either input stimulus. Massaro (1987), through many empirical studies, has shown that the McGurk effect can be accounted for within a broader pattern processing perspective. On Massaro's scheme, the effects of vision on audition reflect Bayesian rules for the combination of auditory and visual inputs, working at the level of phoneme identification information. That is, McGurk stimuli, while producing illusory identifications, nevertheless behave systematically with respect to *all* combinations of possible visual and auditory syllables to which the perceiver is exposed in the experiment. Identical principles apply to the combination of, for instance, heard and written syllables.

[2] In itself, this finding militates against a completely motoric theory of audio-visual speech perception, for at the age of six months the child has no useful speech production abilities.

REFERENCES

Alsius, A., Navarra, J., Campbell, R. & Soto-Faraco, S. S. 2005 Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* **15**, 839–843. (doi:10.1016/j.cub.2005.03.046)

Andersson, U. & Lidestam, B. 2005 Bottom-up driven speechreading in a speechreading expert: the case of AA (JK023). *Ear Hear.* **26**, 214–224. (doi:10.1097/00003446-200504000-00008)

Auer Jr, E. T. & Bernstein, L. E. 1997 Speechreading and the structure of the lexicon: computationally modelling the effects of reduced phonetic distinctiveness on lexical uniqueness. *J. Acoust. Soc. Am.* **102**, 3704–3710. (doi:10.1121/1.420402)

Bernstein, L. E., Auer, E. T., Moore, J. K., Ponton, C. W., Don, M. & Singh, M. 2002 Visual speech perception without primary auditory cortex activation. *Neuroreport* **13**, 311–315. (doi:10.1097/00001756-200203040-00013)

Bernstein, L. E., Auer Jr, E. T. & Moore, J. K. 2004*a* Audiovisual speech binding: convergence or association? In *The handbook of multisensory perception* (eds G. A. Calvert, C. Spence & B. E. Stein), pp. 203–224. Cambridge, MA: MIT Press.

Bernstein, L. E., Auer, E. T. & Takayanagi, S. 2004*b* Auditory speech detection in noise enhanced by lipreading. *Speech Commun.* **44**, 5–18. (doi:10.1016/j.specom.2004.10.011)

Brass, M. & Heyes, C. 2005 Imitation: is cognitive neuroscience solving the correspondence problem? *Trends Cogn. Sci.* **9**, 489–495. (doi:10.1016/j.tics.2005.08.007)

Buccino, G. *et al.* 2001 Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *Eur. J. Neurosci.* **13**, 400–404. (doi:10.1046/j.1460-9568.2001.01385.x)

Burnham, D. & Dodd, B. 2004 Auditory–visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Dev. Psychobiol.* **45**, 204–220. (doi:10.1002/dev.20032)

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M. & Vatikiotis-Bateson, E. 2004 Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *J. Cogn. Neurosci.* **16**, 805–816. (doi:10.1162/089892904970771)

Calvert, G. A. & Campbell, R. 2003 Reading speech from still and moving faces: the neural substrates of seen speech. *J. Cognit. Neurosci.* **15**, 57–70. (doi:10.1162/089892903321107828)

Calvert, G. A. & Lewis, J. W. 2004 Hemodynamic studies of audiovisual interaction. In *The handbook of multisensory perception* (eds G. A. Calvert, C. Spence & B. E. Stein), pp. 483–502. Cambridge, MA: MIT Press.

Calvert, G. A., Bullmore, E., Brammer, M. J., Campbell, R., Woodruff, P., McGuire, P., Williams, S., Iversen, S. D. & David, A. S. 1997 Activation of auditory cortex during silent speechreading. *Science* **276**, 593–596. (doi:10.1126/science.276.5312.593)

Calvert, G. A., Campbell, R. & Brammer, M. 2000 Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol.* **10**, 649–657. (doi:10.1016/S0960-9822(00)00513-3)

Campbell, R., Zihl, J., Massaro, D. W., Munhall, K. & Cohen, M. M. 1997 Speechreading in the akinetopsic patient. *Brain* **121**, 1794–1803.

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). *Cognit. Brain Res.* **12**, 233–243. (doi:10.1016/S0926-6410(01)00054-4)

Capek, C. M., Bavelier, D., Corina, D., Newman, A. J., Jezzard, P. & Neville, H. J. 2004 The cortical organization of audio-visual sentence comprehension: an fMRI study at 4 Tesla. *Cognit. Brain Res.* **20**, 111–119. (doi:10.1016/j.cogbrainres.2003.10.014)

Capek, C. M., Campbell, R., MacSweeney, M., Woll, B., Seal, M., Waters, D., Davis, A. S., McGuire, P. K. & Brammer, M. J. 2005 The organization of speechreading as a function of attention. Cognitive Neuroscience Society Annual Meeting, poster presentation, San Francisco, CA: Cognitive Neuroscience Society.

Capek, C. *et al.* In preparation. Cortical correlates of the processing of stilled speech images—effects of attention to task.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F. & Deltenre, P. 2002 Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. *Clin. Neurophysiol.* **113**, 495–506. (doi:10.1016/S1388-2457(02)00024-X)

Diehl, R. L. 2008 Acoustic and auditory phonetics: the adaptive design of speech sound systems. *Phil. Trans. R. Soc. B* **363**, 965–978. (doi:10.1098/rstb.2007.2153)

Driver, J. 1996 Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature* **381**, 66–68. (doi:10.1038/381066a0)

Fant, G. 1960 *Acoustic theory of speech production*. The Hague, The Netherlands: Mouton.

Fowler, C. A. & Dekle, D. 1991 Listening with eye and hand: crossmodal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* **17**, 816–828. (doi:10.1037/0096-1523.17.3.816)

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L. & Logothetis, N. K. 2005 Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J. Neurosci.* **25**, 5004–5012. (doi:10.1523/JNEUROSCI.0799-05.2005)

Grant, K. W., Greenberg, S., Poeppel, D. & van Wassenhove, V. 2004 Effects of spectro-temporal asynchrony in auditory and auditory–visual speech processing. *Semin. Hear.* **25**, 241–255. (doi:10.1055/s-2004-832858)

Green, K. P., Kuhl, P. K., Meltzoff, A. N. & Stevens, E. B. 1991 Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Percept. Psychophys.* **50**, 524–536.

Hall, D. A., Fussell, C. & Summerfield, A. Q. 2005 Reading fluent speech from talking faces: typical brain networks and individual differences. *J. Cogn. Neurosci.* **17**, 939–953. (doi:10.1162/0898929054021175)

Hickok, G. & Poeppel, D. 2004 Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* **92**, 67–99. (doi:10.1016/j.cognition.2003.10.011)

Jordan, T. R. & Sergeant, P. C. 2000 Effects of distance on visual and audiovisual speech recognition. *Lang. Speech* **43**, 107–124.

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. *Br. J. Audiol.* **34**, 225–230.

MacSweeney, M. *et al.* 2002 Neural systems underlying British Sign Language and audio-visual English processing in native users. *Brain* **125**, 1583–1593. (doi:10.1093/brain/awf153)

Massaro, D. W. 1987 *Speech perception by ear and by eye*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McGurk, H. & MacDonald, J. 1976 Hearing lips and seeing voices. *Nature* **264**, 746–748. (doi:10.1038/264746a0)

Meltzoff, A. N. & Moore, M. K. 1983 Newborn infants imitate adult facial gestures. *Child Dev.* **54**, 702–709. (doi:10.2307/1130058)

Miller, L. M. & D'Esposito, M. D. 2005 Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J. Neurosci.* **25**, 5884–5893. (doi:10.1523/JNEUROSCI.0896-05.2005)

Möttönen, R., Schurmann, M. & Sams, M. 2004 Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. *Neurosci. Lett.* **363**, 112–115. (doi:10.1016/j.neulet.2004.03.076)

Möttönen, R., Järveläinen, J., Sams, M. & Hari, R. 2005 Viewing speech modulates activity in the left S1 mouth cortex. *Neuroimage* **24**, 731–737. (doi:10.1016/j.neuroimage.2004.10.011)

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T. & Vatikiotis-Bateson, E. 2004 Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychol. Sci.* **15**, 133–137. (doi:10.1111/j.0963-7214.2004.01502010.x)

Nishitani, N. & Hari, R. 2002 Viewing lip forms: cortical dynamics. *Neuron* **36**, 1211–1220. (doi:10.1016/S0896-6273(02)01089-9)

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T. & Sams, M. 2005 Processing of audiovisual speech in Broca's area. *Neuroimage* **25**, 333–338. (doi:10.1016/j.neuroimage.2004.12.001)

Patterson, R. D. & Johnsrude, I. S. 2008 Functional imaging of the auditory processing applied to speech sounds. *Phil. Trans. R. Soc. B* **363**, 1023–1035. (doi:10.1098/rstb.2007.2157)

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A. & Sams, M. 2005 Primary auditory cortex activation by visual speech: an fMRI study at 3T. *Neuroreport* **16**, 125–128. (doi:10.1097/00001756-200502080-00010)

Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jäskeläinen, I. P., Kujala, T. & Sams, M. 2006 Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: an fMRI study at 3T. *Neuroimage* **29**, 797–807. (doi:10.1016/j.neuroimage.2005.09.069)

Puce, A., Syngeniotis, A., Thompson, J. C., Abbott, D. F., Wheaton, K. J. & Castiello, U. 2003 The human temporal lobe integrates facial form and motion: evidence from fMRI and ERP studies. *Neuroimage* **19**, 861–869. (doi:10.1016/S1053-8119(03)00189-7)

Radeau, M. & Bertelson, P. 1974 The after-effects of ventriloquism. *Q. J. Exp. Psychol.* **26**, 63–71. (doi:10.1080/14640747408400388)

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

Remez, R. E. 2005 Three puzzles of multimodal speech perception. In *Audiovisual speech* (eds E. Vatikiotis-Bateson, G. Bailly & P. Perrier), pp. 12–19. Cambridge, MA: MIT Press.

Rizzolatti, G. & Arbib, M. A. 1998 Language within our grasp. *Trends Neurosci.* **21**, 188–194. (doi:10.1016/S0166-2236(98)01260-0)

Rosenblum, L. D. & Saldaña, H. M. 1996 An audiovisual test of kinematic primitives for visual speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* **22**, 318–331. (doi:10.1037/0096-1523.22.2.318)

Rosenblum, L. D., Johnson, J. A. & Saldaña, H. M. 1996 Point-light facial displays enhance comprehension of speech in noise. *J. Speech Hear. Res.* **39**, 1159–1170.

Rosenblum, L. D., Schmuckler, M. A. & Johnson, J. A. 1997 The McGurk effect in infants. *Percept. Psychophys.* **59**, 347–357.

Sadato, N. *et al.* 2005 Cross modal integration and changes revealed in lipmovement, random-dot motion and sign languages in the hearing and deaf. *Cereb. Cortex* **15**, 1113–1122. (doi:10.1093/cercor/bhh210)

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. *Neurosci. Lett.* **127**, 141–145. (doi:10.1016/0304-3940(91)90914-F)

Sams, M., Mottonen, R. & Sihvonen, T. 2005 Seeing and hearing others and oneself talk. *Cogn. Brain Res.* **23**, 429–435. (doi:10.1016/j.cogbrainres.2004.11.006)

Santi, A., Servos, P., Vatikiotis-Bateson, E., Kuratate, T. & Munhall, K. 2003 Perceiving biological motion: dissociating visible speech from walking. *J. Cogn. Neurosci.* **15**, 800–809. (doi:10.1162/089892903322370726)

Scott, S. K. 2005 Auditory processing—speech, space and auditory objects. *Curr. Opin. Neurobiol.* **15**, 197–201. (doi:10.1016/j.conb.2005.03.009)

Skipper, J. I., Nusbaum, H. C. & Small, S. L. 2005 Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* **25**, 76–89. (doi:10.1016/j.neuroimage.2004.11.006)

Soto-Faraco, S., Navarra, J. & Alsius, A. 2004 Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition* **92**, B13–B23. (doi:10.1016/j.cognition.2003.10.005)

Sumby, W. H. & Pollack, I. 1954 Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* **26**, 212–215. (doi:10.1121/1.1907309)

Summerfield, A. Q. 1979 The use of visual information in phonetic perception. *Phonetica* **36**, 314–331.

Summerfield, A. Q. 1987 Some preliminaries to a theory of audiovisual speech processing. In *Hearing by eye* (eds B. Dodd & R. Campbell), pp. 58–82. Hove, UK: Erlbaum Associates.

Thompson-Schill, S. L. 2005 Dissecting the language organ: a new look at the role of Broca's area in language processing. In *Twenty-first century psycholinguistics: four cornerstones* (ed. A. Cutler), pp. 173–189. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tiippana, K., Andersen, T. S. & Sams, M. 2004 Visual attention modulates audiovisual speech perception. *Eur. J. Cogn. Psychol.* **16**, 457–472. (doi:10.1080/09541440340000268)

van Wassenhove, V., Grant, K. W. & Poeppel, D. 2005 Visual speech speeds up the neural processing of auditory speech. *Proc. Natl Acad. Sci. USA* **102**, 1181–1186. (doi:10.1073/pnas.0408949102)

Watkins, K. E., Strafella, A. P. & Paus, T. 2003 Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia* **41**, 989–994. (doi:10.1016/S0028-3932(02)00316-0)

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J. & McCarthy, G. 2003 Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb. Cortex* **13**, 1034–1043. (doi:10.1093/cercor/13.10.1034)

Yehia, H. C., Kuratate, T. & Vatikiotis-Bateson, E. 2002 Linking facial animation, head motion and speech acoustics. *J. Phonet.* **30**, 555–568. (doi:10.1006/jpho.2002.0165)