Seeing speech in unexpected places

Mouths, machines and minds

RUTH CAMPBELL *delivered the C.S. Myers Lecture at the Society's Annual Conference in Belfast, April 1999.*

THIS article is about what is going on when we watch faces talking — how we perceive the facial actions produced in speech. Why should psychologists be interested in this? There are two reasons.

Firstly, seeing facial actions improves speech understanding. The most obvious example of this is in noisy conditions when we are trying to follow a conversation – at a party or in a noisy café, for instance. Under these conditions, seeing the speaker can generate a gain in speech understanding equivalent to increasing the auditory signal-to-noise ratio by up to 20 decibels (Sumby & Pollack, 1954).

Computational scientists are now attempting to develop programs that not only convert text to speech, but convert text to a (virtual) seen speaker in action. Should this succeed, the virtual speaker should also be very helpful in noisy environments such as the cockpit of a fighter plane, where the pilot is bombarded with acoustic signals while trying to follow auditory instructions or take in speech-based information. And this could give the inventors an edge in a wide range of media simulations.

Figure 1. In the McGurk effect, viewers see and hear synchronised dubbed incongruent speech sounds, leading to the impression that a different sound was heard.

Secondly, although very few of us are good lipreaders, seeing speech is one of the things that we all do better than we think we do. The most convincing example of this is the 'McGurk' effect (McGurk & MacDonald, 1976), where we think we have *heard* a speech sound (phoneme) that is actually delivered 'by eye'. This can occur when we hear and see synchronised dubbed phonemes that are actually incongruent (see Figure 1).

We now know that vision affects the phonetic context that determines phoneme identification (Green & Kuhl, 1989, 1991). Our understanding of speech sounds is not achieved by ear alone. It is both amodal (the phoneme is an *abstract* linguistic entity) and multimodal. The speech processing system will make use of information relevant to distinguishing the meaning-carrying units in whatever modality they are delivered.

This last insight opens the way to intriguing questions. When we 'read' speech, we read it from the face. A good deal is known about the neuropsychology and neurophysiology of face processing — especially the processing of stilled facial images for identification or for expression. We also know a good deal about auditory speech processing, in cortical terms.

But how can the speech analysis system make use of face analysis networks? Indeed, *does* it make use of face-processing systems at all?

A sceptic might say that speechreading just requires that the viewer analyse mouth actions and that mouths are relatively unimportant for face-reading. But this would be too strong.

For one thing, we can see speech without actually seeing the mouth in action – for instance, if we watch someone speak with a handkerchief stretched over the mouth. The deformations of the skin that accompany mouth and jaw actions can be sufficiently systematic to deliver reliable speech-related information. Also, we don't have to maintain the mouth region in foveal vision (i.e. centred on the fovea — the area of clearest vision) to follow speech well.

Moreover, understanding seen speech can be harder when the facial context is deformed by image manipulation (for instance by inverting the face, but leaving the mouth the right way up). Or it can be harder when several speakers, rather than just one, have to be speechread, as when a series of people make short spoken presentations — one after the other. The

Requests for reprints of this article should be addressed to:
Professor Ruth Campbell
Department of Human Communication Science
University College London
London WC1N 1PG
E-mail: r.campbell@ucl.ac.uk

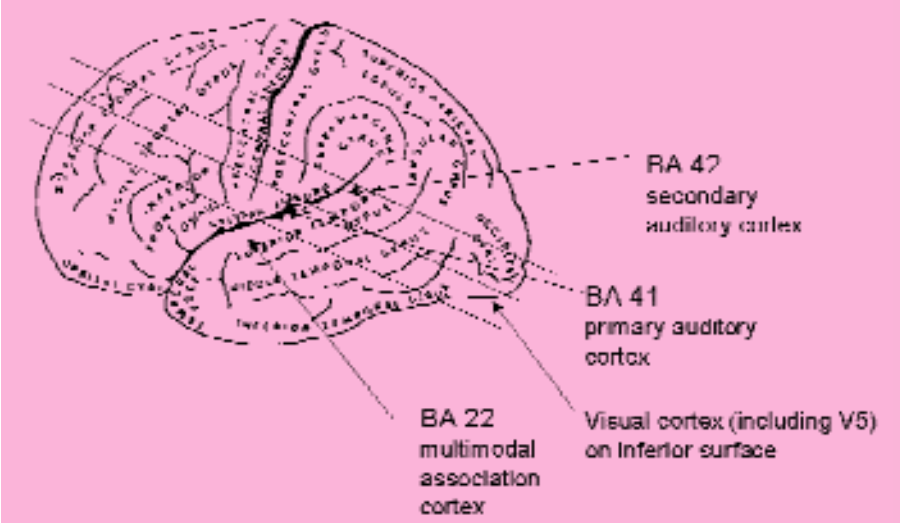

Figure 2. Diagrammatic view of the left surface of the brain, with three sections indicated (see Figure 3 below)

ability to process faces helps to perceive speech from them, although reading faces for speech and for other meanings may dissociate at a later cognitive stage (Campbell *et al.*, 1996*a*).

But there is a sense in which facereading for speech may be special. The identification of a spoken utterance, unlike the identification of an expression of intention or emotion from the face, is dynamically structured, since face actions follow the speech act directly.

To identify a silently spoken 'seven' (a trivial achievement, if the viewer is told 'you'll see a number between one and ten'), it is not sufficient to identify the terminal positions of the face — the dynamic transitions between the states must be processed.

One patient, LM, had lost the ability to perceive visual movement following a cerebro-vascular accident affecting the posterior temporo-occipital cortex (area V5 — see Figure 2). She was unable to identify speechread numbers — although she could identify speech patterns shown as simple stilled images such as 'oo' or 'ff' (Campbell *et al.*, 1997).

Similar findings have emerged with different techniques, such as displays of point-lights on the facial surface (Rosenblum & Saldaña,1998), where the display is no more than a set of moving light dots. These lack visual structure, but maintain the movement characteristics of the face surface when speaking. Such point-light displays can usefully improve the understanding of speech in noise and can even sometimes generate McGurk effects.

The dynamic structure of speechreading, then, may make it special in terms of the neurophysiological mechanisms that support it.

Despite this, speechreading, unlike listening to speech, can be informative at an 'instantaneous' level. We cannot reliably identify an auditory speech event such as a stop-consonant (/b/ or /d/) by isolating and displaying the few milliseconds of the speech spectrogram containing the relevant intensities. By contrast, as we see in Figure 1, some speech *can* be identified from a seen image, and LM was perfectly able to do this.

Stilled speech images can even work as 'McGurk' inducers, affecting the perception of a co-occurring auditory speech event. Vision and audition deliver differently structured percepts; this makes understanding speechreading problematic, yet potentially enlightening.

Neurophysiology of speech, face and perceived movement

What are the cortical bases for processing speech, faces and visual movement? Can such knowledge predict the networks engaged by seeing faces speak ? The 'speech-and-language cortex' lies on the lateral surface of the brain, around the major lateral fold in the brain, the Sylvian fissure, running up to the supramarginal gyrus (see Figure 2).

The superior temporal gyrus, being a long ridge lying along the superior surface of the temporal lobe, following the Sylvian fissure for most of its length up to the supramarginal gyrus, is therefore an important part of the speech and language cortex.

Nestling near the midregion of the superior temporal gyrus, adjacent to an underlying cortical region called the insula, is a very small area termed the primary auditory cortex (also called Brodmann area (BA) 41). This is the first cortical projection site of the auditory nerve. Patients who have bilateral damage to BA 41 are 'cortically deaf'; they lose perception of *any* sound.

In an adjacent area, stretching up the superior temporal gyrus towards the supramarginal gyrus, is the secondary auditory cortex (BA 42). Damage to this region can produce a number of impairments, but the most significant is a loss of the ability to hear speech sounds accurately. That is, phonological perception can be impaired following damage here.

In cortical imaging studies of normal people listening to speech, BA 42 is part of a network which becomes active, bilaterally, when participants perform phonological tasks when reading (e.g. judging whether words rhyme or not).

The lower part of the superior temporal gyrus, known as BA 22, can also be considered part of the auditory cortex, but it has multimodal functions too — especially in relation to perceived visual movement. Interestingly, it may support a linguistic function for sign-language. It can also be activated by other types of biological

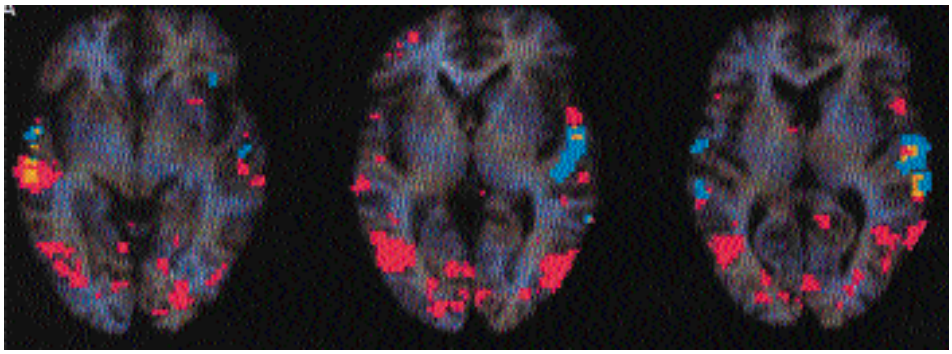

Figure 3. Three consecutive scans (top = front of head) showing areas activated by hearing spoken numbers without seeing the speaker (blue), seeing numbers spoken silently (red) and areas of overlap of activation (yellow). These include the primary and secondary auditory cortices bilaterally.

movement, such as movement of the eyes or mouth (Puce *et al.*, 1998) or of a body in motion (Howard *et al.*, 1996). BA 22 receives projections from the visual cortex, especially from inferior temporo-occipital regions including V5, the visual movement cortex.

These areas (BA 22, 41, 42) are not generally active in viewing still faces — for instance in detecting faces among other types of stimuli. The fusiform gyrus of the inferior temporo-occipital lobe is an important site for face detection, while parahippocampal and inferior temporal regions may be involved in identifying (recognising) faces.

Against this background, then, we can ask — which cortical areas are active in silent speechreading? Calvert *et al.* (1997) were the first to ask this question, and Figure 2 also shows the approximate location of the brain 'slices' that would help to answer it.

Using fMRI (functional magnetic resonance imaging) to show slices of the brain, normal right-handed people were scanned under two conditions: listening to numbers without seeing a face and watching numbers being spoken silently. The baseline for the speechreading condition was watching a still face. Figure 3 shows the patterns of activation for heard, compared with seen, speech corresponding to the three sections indicated in Figure 2.

Hearing activated the primary auditory cortex (BA 41) and parts of BA 42; the control condition was a silent, blank display. Silent speechreading, compared with viewing a still face, activated all of the superior temporal gyrus (STG) and V5. The former included phonological areas (BA 42) and the primary auditory cortex (BA 41), as well as BA 22.

To what extent is STG activation specific to speechreading, and to what extent a function of viewing a face performing any movement? A second experiment explored the activation pattern when watching a face making rhythmic movements at the same rate as speech — but with no mouth opening (gurning) — and 'nonsense' mouth movements with mouth openings that looked 'speechlike' but which were not English (pseudospeech).

Gurning faces activated BA 22, but did not activate BA 42. Pseudospeech activation was indistinguishable from that for speechreading numbers.

These findings taken together confirm that, as well as activating networks involved in identifying faces in action (BA

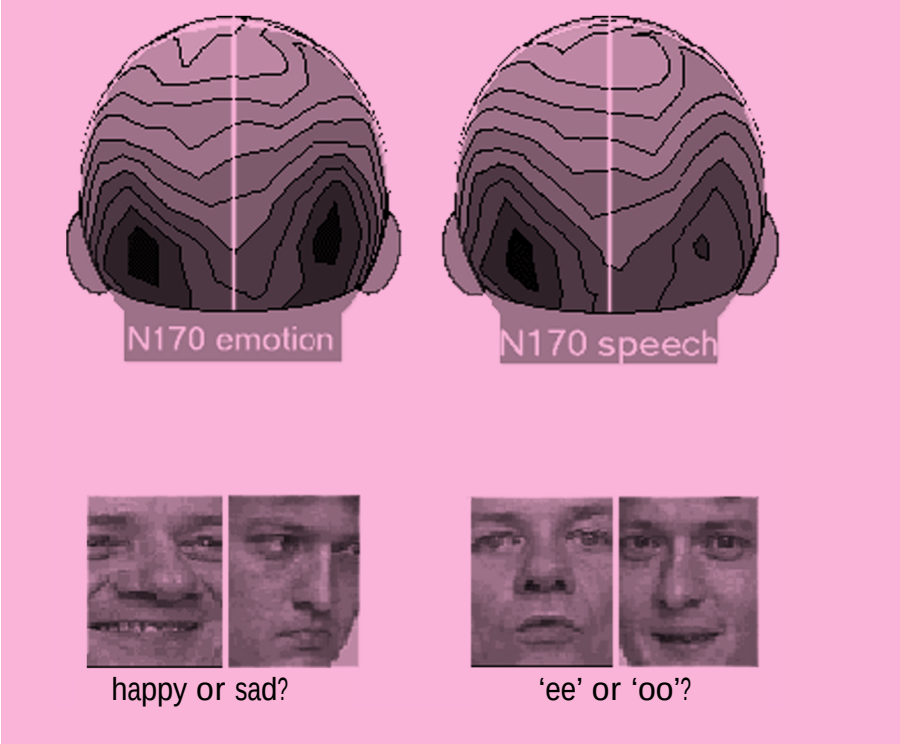

Figure 4. Voltage maps showing scalp potentials for the N170 waveform, for a speeded expression task (top left) and for a speeded speech discrimination task (top right), and samples of the images used to elicit these (bottom)

22), speechreading activates phonological areas of the auditory cortex (BA 42) and the primary auditory cortex (BA 41). To do this, it is not necessary for lexical identification to be achieved — *potential* speech (pseudospeech) as well as real speech produced the same pattern.

The finding that silent speechreading can activate the primary auditory cortex was important; this was the first time that a non-acoustic event had been shown to activate it. This did fit some functional descriptions of speechreading — for example, the finding that seeing silent speech can produce a functional 'echo' in immediate recall (Campbell & Dodd, 1980).

When one watches someone silently speak a list of numbers, the last one seems to persist in immediate memory, as if it leaves an echo. This happens for heard speech, too, but not for written lists. However, the finding that the primary auditory cortex could be activated by silent speechreading casts doubt on the idea that the sensory cortex is immutably tuned to a specific modality.

Now, the fMRI scanner is very noisy in operation, producing a loud (approximately 60 decibels) rhythmic white noise for every scan performed. Could activation of the primary auditory cortex by silent speechreading reflect what occurs when the speechreading task is done in the presence of scanner noise?

MacSweeney *et al.* (in press) have recently been exploring this, using an event-related analysis. fMRI measures cortical bloodflow patterns, and these can lag neural activity by up to five seconds. Thus, when the scan is performed five or so seconds after a silent speechreading task has ended, the pattern of activity can be related to the earlier task without noise affecting the pattern. Under these conditions, findings to date suggest that all the auditory cortex is active, just as in the first studies.

These findings do not necessarily mean that seeing speech accesses the primary auditory cortex directly. fMRI, like other cortical imaging techniques such as PET (positron emission tomography), is extremely informative about where in the brain an activity occurs, but is less informative about the time-course of these events. Using these techniques we can only uncover the sequence of processing operations by indirect means — for example, by constructing tasks designed to investigate early and late processes separately.

There are, however, online indicators which measure the electromagnetic potentials that accompany neural events. These include event-related scalp measurements — measuring changes,

either in the orientation of the magnetic field or in the electrical potential (ERP — event-related potential), that accompany cognitive events.

Listening passively to spoken syllables (no vision) while in the magnetometer can reliably induce a waveform called the mismatch negativity (MMN). This occurs about 100 milliseconds after stimulus onset, and the waveform is largest over the temporal areas of the scalp which lie directly above the auditory cortex.

Sams *et al*. (1991) showed that the MMN can be modified when a visual syllable is dubbed to the auditory one — that is when a McGurk stimulus is projected. The integration response produced by seeing the speaker occurs later than the auditory MMN (by about 200 milliseconds), but it is similarly localised. It is greatest over the temporal parts of the scalp.

One face-related waveform has been identified over posterior parts of the scalp, with a source in the visual areas of the brain. The N170 is a waveform generated when a face — or sometimes a face part — is presented. It is specific to faces, in that it is of larger amplitude and occurs earlier for faces than for any other type of picture.

The source of the N170 waveform may be the fusiform gyrus, corroborating cortical imaging findings for face detection. To date, however, this waveform has not been shown to be sensitive to specific face tasks — it appears under all sorts of viewing conditions (Bentin *et al*., 1996). Michelle De Haan and I have been trying to find whether there is an electrophysiological signature for viewing a speaking face, using stilled face images.

Figure 4 shows the summarised voltage maps for the N170 waveform, obtained under two different speeded-choice tasks on a single facial image — 'is the face saying "oo" or "ee"?' and 'is the face "happy" or "sad"?'.

In an earlier experimental study, it had been shown that the speech-matching, but not the 'happy–sad' task, was performed faster when the images were projected to the left hemisphere; it was a task that engaged the language-processing hemisphere (Campbell *et al*., 1996*b*). The voltage maps also show a marked left-side advantage for the speech task.

For these types of image at least, some specialisation for processing faces for speech may occur at a relatively early perceptuo-cognitive stage; the N170 occurs within 200 milliseconds of stimulus onset. However, we know this cannot be the whole story. Not only are stilled faces a very special form of 'speech carrier', but other studies, in neuropsychological patients and in normal people, have shown that the right hemisphere certainly can be recruited in speechreading (for a review, see Campbell, 1996).

Conclusions and directions

It is sometimes claimed that speechreading is really a form of problem solving. While this may be true for some forms of expert speechreading, it is not true for the conditions we have tested. Where simple, highly constrained material is speechread, we do not find specific problem-solving or compensatory strategies are used. There is no behavioural or neurophysiological suggestion that this occurs.

Rather, I believe our studies show that seeing speech makes its major impact by directly implicating specialised cortical speech processing areas. Those posterior cortical areas specialised for perceiving movement and for perceiving visual forms must be functionally intact for speech perception areas to be engaged by seeing actions on a speaker's face.

There may be a number of different specialisations that can allow visible speech to do this. These may recruit sites in the two hemispheres differentially depending on the task; neuroimaging does not yet tell us more about this.

Obviously, this is just the start of the story. In ongoing work with colleagues, speechreading in deaf people is being investigated, and the precise cortical systems involved in processing stilled speech images are being contrasted with those for speech in motion.

Current work is also elucidating the cortical bases of audiovisual speech. Could audiovisual speech enhance activity in primary auditory or visual processing areas? This may open the way to new therapeutic initiatives for the hard-of-hearing and possibly for youngsters with problems in auditory speech perception that do not have an obvious sensory base. It may yet redefine notions about the plasticity of sensory processing systems.

Acknowledgements

This article reports collaborative work conducted with Michael Brammer, Gemma Calvert, Tony David, Philip McGuire, Mairead MacSweeney, Bencie Woll; it is now funded by the Medical Research Council. Examples and insights from the work of Eric Bateson, Hani Yehia, Takaaki Kuratate, Mike Cohen and Dominic Massaro were used to illustrate the talk. Michelle De Haan initiated the ERP neurophysiological work reported here. The interpretations offered here are, however, my own.

This article is dedicated to the memory of Christian Benoit, Kerry Green and Harry McGurk, who have advanced the understanding of audiovisual speech, and who died in 1998.

References

Bentin, S., Allison, T., Puce, A., Perez, E. & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, 8, 551–565.

Calvert, G., Bullmore, E., Brammer, M., 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.

Campbell, R. (1996). Seeing brains seeing speech: A review and speculations. In D.G. Stork & M. Henneke (Eds), *Speechreading by Humans and Machines: Models, Systems and Applications* (NATO ASI series). Berlin: Springer.

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

Campbell, R., Brooks, B., De Haan, E.H.F. & Roberts, A. (1996*a*). Dissociated face processing skills: Seen speech, expression and identity matching from photographs: Reaction time evidence. *Quarterly Journal of Experimental Psychology*, 48/49A, 295–332.

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

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

Green, K.P. & Kuhl, P.K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception and Psychophysics*, 45, 34–42.

Green, K.P. & Kuhl, P.K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 278–288.

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

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

MacSweeney, M., Amaro, E., Calvert, G., Campbell, R., David, A.S., McGuire, P., Williams, S., Woll, B. & Brammer, M. (in press). Activation of auditory cortex by silent speechreading does not require scanner noise: An event-related fMRI study. *Human Brain Mapping*.

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.

Rosenblum, L.D. & Saldaña, H.M. (1998). Time-varying information for visual speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds), *Hearing By Eye II*. Hove: Psychology Press.

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

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