

SEEING SPEECH IN SPACE AND TIME PSYCHOLOGICAL AND NEUROLOGICAL FINDING

**Ruth CAMPBELL-
Londres**

**Department of Human Communication Science,
University College London, Grande Bretagne**

Hearing by eye : speechreading as a cognitive process.

While speechreading -viewing the movements of the speaker's lower face -cannot deliver all the constrastive information of heard speech, it can certainly help to understand speech in noise or in conditions of impaired hearing. In particular, seeing the speaker can sometimes help to deliver the phonetic features that cannot be heard (e.g place of articulation, can be seen by lip position and mouth action).

How does the visual system manage this type of task ? Studies with brain-lesioned patients can indicate the critical cortical bases for effective speechreading, and studies with normal hearer-viewers confirm and extend these findings. Brain imaging studies offer a newer perspective on these issues. This talk will address two main questions:

- (1) what is the relative importance of resolving visual movement and of resolving visual shape in speechreading ?
 - (2) how does reading speech from faces differ, how resemble, reading other facial actions (such as expression) ?
- The answers indicate just how speechreading might be special.

Entendre grâce aux yeux : la lecture labiale comme processus cognitif

La lecture labiofaciale, c'est-à-dire l'observation du bas du visage de la personne qui parle -même si elle ne permet pas autant que l'ouïe de saisir les nuances de ce qui est dit, aide à comprendre le sens de la parole dans un environnement bruyant ou si l'on est malentendant. Le fait de voir la personne qui parle aide parfois à déceler les nuances phonétiques non perçues par l'ouïe (mouvement des lèvres et articulation par la bouche). Comment le système visuel peut-il remplir cette fonction ? Des observations effectuées avec des personnes atteintes de lésions cérébrales ont permis de déceler le rôle critique du cortex dans la lecture labiale, comme cela a pu être constaté chez des personnes entendant. A cet égard, la technique de l'imagerie cérébrale ouvre de nouvelles perspectives. Nous nous proposons d'aborder deux questions essentielles :

- 1) quelle est l'importance relative de l'interprétation donnée par le regard d'une part aux mouvements et d'autre part à la forme du visage de la personne qui parle ?
- 2) dans la lecture labiofaciale quels sont les rôles respectifs de l'observation du visage et d'autres actions faciales (telles que l'expression) ? La réponse à cette question révèle le caractère très particulier de la lecture labiale.

SEEING SPEECH IN SPACE AND TIME : PSYCHOLOGICAL AND NEUROLOGICAL FINDINGS ABSTRACT

Unlike heard speech, some aspects of seen speech can be processed independently of their dynamic (temporal) characteristics. The extent to which speechreading may be supported, separately, by visual mechanisms for the analysis of still and moving events is considered in relation to some experimental findings with normal

speaker/viewers, to results with neurological patients with circumscribed lesions to parts of the visual system that support the perception of movement, and to cortical imaging studies.

I-INTRODUCTION

It seems obvious that seeing speech involves both the analysis of visual form and of the dynamic characteristics of the seen articulators. Both mouthshape (and the visibility of mouth parts) and mouth movement (the dynamics of mouth actions, including rate of speech) play their part. This insight is confirmed by a range of experimental findings^{20,22,24} and by findings in applied telematics which show that speechreading accuracy for audiovisual inputs with auditory dynamic noise falls off as frame-rate (temporal resolution) of the display of the speaker's face drops from about 30 Hz to 8-12 Hz^{24, 25} in addition, seen rate of speech can be readily discriminated and can directly affect the identification of a heardspeech token^{6, 7}. But how does this work ? Is one process subsumed in the other ? Do we infer form from motion or decompose form into movement trajectories ? Are they integrated or separated in processing ? How crucial is perceived motion to the perception of speech – to what extent can it deliver speech information when visual form cannot ? In this paper I present some experimental data from normal hearing subjects, evidence from neurological patients with focal lesions to visual cortex and some suggestive cortical imaging data. All of this suggests that the contributions of movement and form to speechreading are separable and have differing roles to play in effective speechreading.

II- EXPERIMENTAL STUDIES OF STILL AND MOVING FACES

Rosenblum and colleagues¹⁶ have shown that point-light displays of a moving face, which cannot be construed as a face in still-frame, can nevertheless affect the report of auditory tokens synchronised to the display. In these studies, for example, seen "va" and heard "th-" could give rise to the auditory perception of "v-". These displays can also aid speechreading in noise, giving an appreciable gain when shadowing noisy speech over audition alone¹⁷. Since such displays deliver minimal information about the visual form of their source, such demonstrations, like those for point-light displays of whole body actions¹⁰, show that perceived movement can contribute to speechreading. However, these effects, while significant, were not as great as when the natural face was viewed. Pointlight displays show that visual movement is important, but do they say more ? It would be tempting to suppose that the point-light displays work by delivering movement trajectories of the (illuminated) parts of the face that articulate speech. Just such facial action patterns can be identified from optic flow patterns in the computational simulation of visible speech¹¹. But are such algorithms psychologically real ? There could be sufficient redundancy in the sparse dot displays to afford (schematic) information about the relative visual forms : -the mouthshape of the vowel, the position of the teeth and lips (teeth were illuminated). Another way to explore the extent to which movement might support speechreading is to use a display where visual form is better specified, but which cannot be speechread, and to see whether animation then delivers speechreading.

A brightness inverted (photographic negative) face in action can maintain all movement information of the speaking face without any corresponding loss in visual form. Yet faces in negative are notoriously hard to identify or to read. In our study we used a bearded speaker, making resolution of facial features even more problematic. Our speaker was videorecorded saying "ba, ga, da and ha" and the resultant images and sound were computer-grabbed (Apple System-7). Using commercial software (Adobe photoshop) image and speech samples were recombined to produce congruent and incongruent tokens (i.e. seen and heard "ba" ; seen "ba" and heard "ga"). These were entered into experimental display software (Apple Superlab) and twenty undergraduate English-speaking subjects, with normal corrected vision and hearing were tested.

When the negative display was shown silently (video alone condition) subjects were at chance at detecting whether the face was saying "ba" or "ga". By contrast, performance was at ceiling for all subjects when the normal (positive) image was seen. When sound was added to the negative display (30 trials on different combinations of tokens) there was **no** influence of vision on audition, that is, there were no McGurk effects. The McGurk effect is when "da" is

"heard" by the subject when "ga" is seen synchronised to a heard (dubbed) "ba"¹². In the corresponding positive viewing condition, McGurk illusions were in the region of 15-20 %.

Thus, where it is almost impossible to speechread the face because of brightness inversion of the image, movement cannot recover speechforms. Unlike point-light displays, such stimuli maintain all the relevant image properties of a face. If movement were critically involved in discriminating spoken /b/ from /d/ we would have expected subjects to recover this information from the dynamics of the display, since all the face-feature landmarks that could serve as point trajectory sources are available in both the negative and the positive display.

Using **positive** images of the same speaker, a series of further explorations confirmed that natural movement was not necessarily useful in recovering speech information. We orthogonally varied temporal and spatial information in the display by changing the frame rate and by adding (Gaussian) blur to the image in a systematic fashion. It has been reported that McGurk effects cannot be found when a still image and a spoken sound are combined¹⁶. But event perception (event parsing) is almost certain to play a role here, leading the subject to attend to the still face as one event, and the heard speech as another. At very frame rates (~20 Hz) frame-rate for incidence of auditory-visual fusions. In the slow-rate condition, just three facial images were used - an open mouth ("-ah") a closed mouth ("b") and a jaw-drop ("ha") The sequence open-closed-jawdrop therefore showed "ab-ha" in animation, while jawdrop-close-jawdrop was perceived as "aha" or "a-da"¹⁸. These were dubbed by hand to the relevant auditory tokens, matching perceptual onset for heard and seen conditions : both 20 Hz and 2 Hz images gave 10-20 McGurk reports. By contrast, all McGurk sensitivity was lost when the images were degraded by visual blur.

In the 2 Hz condition the images did not have normal dynamic characteristics, they appeared "jumpy" and unnatural. Nevertheless, McGurk illusions occurred. Natural movement in itself may be less crucial to the identification of seen speechsounds -at least of the type that give rise to these segmental illusions - than the delivery of sampled ingoemsation about changes in mouth and lipshape. McGurk effects and possibly other speechreading percepts may depend **neither** on fully temporally specified visual speech **nor** on fully spatially (form) specified visual speech. Such effects are robust and reflect "natural" ecological contingencies and cross-modal information redundancies.

a The experiments with positive-face animations were inspired by an undergraduate project of Fiona Christie (University of Durham). She and Vicki Bruce are thanked for their comments at various times on aspects of this work.

III- NEUROPSYCHOLOGICAL EVIDENCE

Within the primate visual system, specialised processing channels are recruited for the perception of motion and of form, and project in distinct ways to primary cortical sites in striate (V1, V2) and extrastriate occipital (V3, V4) and occipitotemporal (V5) cortex. Areas V1 and V2 are the primary occipital sites : damage here impacts on perception at higher levels. Speechreading has been explored in a number of patients with circumscribed lesions to these areas. It includes the identification of speech sounds from photographs of mouth-shapes, and susceptibility to McGurk illusions. These are abilities shown by normally hearing people whatever their skill at interpreting silently spoken speech in context.

- Patient HJA⁹ had some damage to V1, V2 and, importantly, to V4, the "form-from-color" area. He was unable to classify photographs of faces in terms of their speechsounds, but could speechread moving faces normally². In HJA, area V5, a specialised site for the perception of visual movement, was undamaged.
- Patient WM, who has intact motion perception, but who also has more extensive lesions to V1, V2 and V4 than HJA, could not speechread at all^{15,23} Movement may help speechreading when some form vision is intact, but cannot replace it when it is absent.
- Patient DF, with individually small but very extensive lesions to V1 and V2 makes a similar point : When tested by direct visual confrontation ("what is this?") this patient could not identify pictured objects reliably⁸. She has been thoroughly investigated for her ability to use visual information to guide action despite this loss of phenomenal perception¹⁴. V5 and parts of V3 are spared and she can report directional movement in random-dot displays. But she is unable to speechread, showing no effect of vision on audition when congruent and incongruent tokens are presented, or when visual tokens ("ba", "va", "da", "tha") are presented alone⁵.

- Patient LM has lesions to V5 bilaterally, with spared VI, V2, V3 and V4¹⁹ : that is she has a functional lesion that is the opposite of that described for DF. It was due to an aneurysm in a venous malformation which occurred eighteen years ago. She can identify pictures of faces, objects and letters, but she is unable to make effective use of visual movement ²⁷. When tested recently, she still had great difficulty in coping with the moving visual world. Nevertheless she could identify speech patterns from photographs. However in reading natural speech she could usually only identify the initial or the final mouthshape and was insensitive to seen rate of speech. She showed no influence of vision on audition when "ba" and "ga" -type tokens were shown which generated McGurk effects in most viewers ⁴.

These patients show us that form and movement in seen speech may be processed - at least initially - by separable systems. There is no "privileged route" to speechreading : for instance through covert imitation of the speechpattern which might afford the emergence of an action-based percept. This is one possibility that might have allowed DF to speechread. Nor can speechreading use an apparently spared visual processing route for the perception of biological motion (as in the point-light figures described by Johansson¹⁰). LM shows spared perceptual sensitivity to biological motion for whole body actions ¹³ but not for seen speech. The pattern of LM's motion blindness suggests that for speech the role of the movement perception mechanisms in V5 may be to perform efficient transitions between endpoints such as specific mouthshapes, teeth-visibility-patterns etc. The patients studied to date underline the point emphasized earlier : while seen movement can contribute to speechreading, it may not deliver information **directly**, through translation of dynamic properties of the display.

IV- CORTICAL IMAGING

A project initiated by G. Calvert and A. S. David has been investigating the brain localisation of visual speech using fMRI cortical imaging techniques. fMRI makes use of the interactions between a strong magnetic field (1,5 tesla) and the small electro-magnetic field established by localised brain activity to map that brain activity. As in other cortical imaging methodologies, the favoured experimental paradigm is subtractive : that is, as far as possible, an experimental and a baseline condition are contrasted in one experiment. The assumption is that any pattern of imaged brain activity reflects the differential activation of the experimental over the control condition.

A series of experiments with volunteer right-handed adult subjects has now been completed to investigate the cortical systems activated by speechreading. First, Calvert et al. confirmed that heard speech (listening to spoken numbers between one and ten) activates classical auditory reception areas (especially Brodmann's area 41). This was itself not trivial : the participant is always subject to noise in the scanner and we were unsure if an auditory speech stimulus would indicate further specific areas of activation. Of the further tested stimulus conditions, that for silent speechreading of the same (number) material is of prime interest here. Statistical analysis showed that silent speechreading activated many posterior and superior temporal areas, bilaterally. Of especial note, primary auditory cortex was activated by this silent task, as well as all the perisylvian speech-association areas and posterior occipito-temporal areas, including V5. Large swathes of superior temporal cortex bilaterally (area 22) were activated by silent speechreading, also.

Further studies have confirmed that this pattern of activation is specific to speechreading : it is obtained when watching "plausible speech" - a range of rhythmic mouth actions including mouth openings, but not for closed mouth ("chewing") actions.

The most important inference from these studies is that silently speechreading a moving face activates auditory receptive cortex. In addition, specific superior posterior temporal sites extrinsic to V5 are active bilaterally. Different sites (right parahippocampal gyrus and the right fusiform gyrus) are implicated for viewing faces for identity or for facial expression judgements ¹⁸. It is already clear from a number of experimental studies ³ that, unlike heard speech or face analysis for identification or expression/intention, the pattern of cortical lateralisation for speechreading may depend on the particular task.

V- CONCLUSION

In speechreading we make use both of the visual form afforded by the pattern of the mouth, tongue and teeth and also of the movement characteristics (rate of jawdrop, mouth-shaping) of speaking. Neuropsychological studies with patients who have specific damage to either the visual form or visual movement systems show us that neither, alone, can deliver effective speechreading. This tells against a theory of speechreading as a form of direct visual movement perception : a theory implied by the point-light experiments of Rosenblum and colleagues. Moreover, experimental evidence using systematic manipulations of brightness-relations, of spatial and of temporal frequency of facial images, failed to support the direct movement hypothesis. We have identified a posterior superior temporal site which is active in viewing a moving face speak and which is not active when viewing a still face. That is, there is clear cortical separation of the visual systems that support the analysis of visual speech from its visual form and from its movement characteristics. The extent to which these may be implicated differently in different groups, including dyslexics as well as in deaf people who do and do not speechread is the next question to be explored.

ACKNOWLEDGEMENTS

Work reported here was supported in part by ESRC project grant R000233622 (R. Campbell & E. H. F. de Haan) and a Research Fellowship from the Leverhuime Foundation (1994 - 1995). Barbara Brooks ran some of the experiments. I also thank all colleagues involved in production of test materials, the testing of patients and in cortical imaging, especially D. Massaro, M. M. Cohen, M. Regard, G. Calvert, A. S. David and J. Zihl. Parts of this work were reported to the European Society of Cognitive Psychology (Rome, September 1995).

This talk is substantially similar to one given at the audiovisual speech symposium at the International Congress for Spoken Language Processing, Philadelphia, 1996 (October).

REFERENCES

1. Calvert G., Bullmore E., Brammer M., Campbell R., Woodruff P., McGuire P., Williams S., Iversen S. D. & David A. S. Activation of auditory cortex during silent speechreading (submitted).
2. Campbell R. The Neuropsychology of Lipreading. *Phil Trans Roy Soc, B*, 335, 39-45, 1992.
3. Campbell R. Neurological bases of speechreading. In D. Stork & M. Henneke (Eds) *Speechreading by humans and machines : a NATO-ASI symposium*. Springer, Heidelberg, 1996.
4. Campbell R., Zihl J., Massaro D., Cohen M. M. & Munhall K. Speechreading in a motion-blind patient submitted (a).
5. Campbell R., Massaro D., Cohen M. M., Goodale M., Zihl J. Posterior cortical impairment and speechreading, submitted (b).
6. Green K.P. The perception of speaking rate from a talker's face *Perc'n and P'physics* 6 587-593, 1987.
7. Green K. & Miller J. On the role of visual rate information in phonetic perception *Perc'n and P'physics*, 38 269-276, 1986.
8. Humprey G. K., Goodale M. A., Jakobson L. J. & Servos P. The role of surface information in object recognition : studies of a visual form agnostic and normal subjects *Perception* 23, 1457-1481, 1994.
9. Humphreys G. W., Donnelly N. & Riddoch J. Expression is computed separately for moving and static faces *Neuropsychologia*, 31, 173-181, 1993.
10. Johansson G. Visual perception of biological motion and a model for its analysis *Perc'n and P'physics*, 14, 201-211, 1973.
11. Mace K. & Pentland A. Lipreading by optical flow (in Japanese) *IEICE Transactions J73-D-II*, 6, 796-803, 1990.
12. McGurk H. & MacDonald J. Hearing lips and seeing voices. *Nature*, 264, 746-748, 1976.
13. McLeod P., Zihl J., Perrett D.I., et al. The perception of biological motion in the motion-blind patient *Visual Cognition*, 3, 1996, 371-393.
14. Milner A. D., Perrett D. I. Johnston R. S., et al., Perception and action in visual form agnosia *Brain*, 114, 405-408, 1991.
15. Regard M., Campbell R. & Landis T. Tests of speechreading in neurological patients (unpub'd).
16. Rosenblum L. D. & Saldaña H. M. Visual primitives for audiovisual speech integration *JEP HP & P*, 22, 1996, 318-331.

17. Rosenblum L. D., Johnson J. A. & Saldana H. M. Visual kinematic information for embellishing speech in noise, JSHR (in press).
18. Sergent J. Brain Imaging Studies of Cognitive Functions *TINS*, 17, 221-227, 1994.

19. Shipp S., de Jong B. M., Zihl J. et al. The brain activity related to residual motion vision in a patient with bilateral lesions of V5 *Brain* 117, 1023-1038, 1994.
20. Summerfield A. Q., McLeod A., McGrath M. & Brooke M. Lips, teeth and the benefits of lipreading. In A. W. Young & H. D. Ellis (Eds) *Handbook of Research in Face Processing*, North Holland, Amsterdam, 218-223, 1989.
21. Summerfield A. Q. Visual perception of phonetic gestures. In I. G. Mattingley & M. Studdert-Kennedy (Eds) *Modularity and the Motor theory of Speech*, Hillsdale, New Jersey, Lawrence Erlbaum, 1991.
22. Summerfield A. Q. Use of Visual information for phonetic processing *Phonetica*, 36, 314-331, 1979.
23. Troscianko T., Davidoff J., Humphreys G. et al. Human Colour Discrimination based on a non-parvocellular pathway *Current Biology*, 6, 1996.
24. Vatakotis-Bateson E. & Munhall K. In D. Stork & M. Henneke (Eds) *Speechreading by Man and machine ; a NATO symposium*, Springer, Heidelberg, 1996.
25. Vitkovitch M. & Barber P. Effect of Videoframerate on shadowing JSHR, 37, 1204-1211, 1995.
26. Vitkovitch M. & Barber P. Visible speech as a function of image quality *Applied Cognitive Psychology*, 10, 121-140, 1996.
27. Zihl J., von Cramon D., Mai N. & Schmid C. Disturbance of movement vision after bilateral posterior brain damage *Brain*, 114, 2235-52, 1991.