SEEING SPEECH IN SPACE AND TIME: PSYCHOLOGICAL AND NEUROLOGICAL FINDINGS

Ruth Campbell

Department of Human Communication Science, University College London, London WC1

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.

1. 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 [18, 20, 22] 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 30Hz to 8-12 Hz [23, 24] In addition, seen rate of speech can be readily discriminated and can directly affect the identification of a heard speech token [5, 6] 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

2.EXPERIMENTAL STUDIES OF STILL AND MOVING FACES

Rosenblum and colleagues[14] 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 alone[15] . Since such displays deliver minimal information about the visual form of their source, such demonstrations, like those for point-light displays of whole body actions[8] , 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 speech[9] . 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 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 'ba' is seen synchronised to a heard (dubbed) 'ga' [10] . In the corresponding condition under normal viewing, McGurk illusions were in the region of 15-20%.

Under these conditions, 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 combined [14]. 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 slow frame rates (say 2-4Hz), while natural movement cannot be seen, the audio-visual event is more likely to be categorised as a single event than when a single, still face frame is seen in relation to a dynamic heard speech sound. We therefore contrasted slow frame-rate (2Hz) and more natural (~20Hz) frame-rate displays 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'[i]. These were dubbed by hand to the relevant auditory tokens, matching perceptual onset for heard and seen tokens. Frame-rate hardly affected the incidence of McGurk fusion illusions under these conditions: both 20Hz 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 information 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 informational redundancies.

3. NEUROPSYCHOLOGICAL EVIDENCE

Within the primate visual system, specialised processing channels are recruited for the perception of motion and of form, and these 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. Speechreading is here taken to be the ability to read speech from the face at any level at all. It includes the identification of speech sounds from photographs of mouth-shapes, and susceptibility to McGurk illusions. These are abilities that normal hearing people have whatever their skill at interpreting silently spoken speech in context.

- Patient HJA[7] 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 normally [1] . 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 [13, 21] . 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 or real objects. She has been thoroughly investigated for her ability to use visual information to guide action despite this loss of phenomenal perception[12]. 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 alone[4].
- Patient LM has lesions to V5 bilaterally, with spared V1,V2,V3 and V4 [17]: 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 [25]. 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 [3].

These patients show us that form and movement in seen speech are 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 Johansson [8]). LM shows spared perceptual sensitivity to biological motion for whole body actions[11] 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.

4. 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. In a series of experiments with right-handed adult subjects, we confirmed, firstly, that heard speech (listening to spoken numbers between one and ten) activates classical auditory reception areas (especially Brodmann's area 41) in the left hemisphere. Secondly, we showed that adding redundant visual displays (congruent visual speech) to heard speech activated the appropriate visual areas (primarily V5). Thirdly, when subjects were speechreading silently spoken numbers (experimental condition) and this was contrasted with audiovisual presentation of the same material (control condition) - some activity of the primary visual areas (V1 and V2) and of some slightly posterior association (temporal) cortex occurred. This suggests that, by and large, both silent speechreading and audio-visual speech activate the same substrate comprising the auditory speech reception areas (Wernicke's area) centring on Brodmann's area 41 in the left hemisphere. There is a slight yet significant additional contribution for seen speech from primary visual areas and from some superior temporal sites, bilaterally.

A fourth condition surprised us. In this, silent speechreading (counting speechread numbers), was contrasted with a baseline condition of viewing a still face, while counting 'internally' during presentation. We had anticipated that left temporal sites would be active including the classical speech reception sites - since we had inferred such activation from the contrast between the audio-visual and visual (speechreading) conditions described above. But what we found was that, as well as expected recruitment of V5 (visual movement) a large **right** hemisphere superior temporal region became active (Brodmann's area 22). This activation was noteworthy for its robustness and reliability: it was the area of strongest activation in all subjects. Identical activation sites were found in a congenitally deaf woman who speechreads.

We infer from these studies that speechreading (a moving face) activates auditory receptive cortex. In addition, a specific, right hemisphere site extrinsic to V5 appears to be active when viewing a moving (speaking) face. Different sites (right parahippocampal gyrus and the right fusiform gyrus) are active when people view pictures of faces for identity or for facial expression judgements [16]. A tantalising possibility is that the right temporal site we have found to be active when silently speechreading is specific to this skill. It may, however, be active for other face movement tasks, such as identifying gestures of expression or identifying someone's face in action. Audio-visual speech makes use of auditory speech areas in the left hemisphere, with additional recruitment of posterior visual areas. It is already clear from a number of experimental studies[2] that, unlike heard speech or face analysis for identification or expression/intention analysis, speechreading need not be localised to one or the other cerebral hemisphere. The particular task processing demands set the pattern of lateralization.

5. 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 right 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 Leverhulme Foundation (1994 -95). Barbara Brooks ran some of the experiments. I also thank all colleagues involved in production of test materials testing 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).

References

1 Campbell,R. The Neuropsychology of Lipreading. *Phil Trans Roy Soc, B, 335*, 39-45,1992

2 Campbell, R Neurological bases of speechreading. In D.Stork & M.Henneke (Eds)*Speechreading by Man and machine: a NATO symposium*, Springer, Berlin, in press

3 Campbell,R., Zihl,J., Massaro,D., Cohen,M.M. & Munhall,K. Speechreading in a motion-blind patient submitted (a)

4 Campbell, R., Massaro,D., Cohen,M.M., Goodale,M., Zihl J.Posterior cortical impairment and speechreading, submitted (b)

5 Green,K.P. The perception of speaking rate from a talker's face *Perc'n and P'physics 6* 587-593, 1987

6 Green,K. & Miller,J. On the role of visual rate information in phonetic perception *Percep'n and P'physics, 38* 269-276, 1986

7 Humphreys,G.W., Donnelly,N. & Riddoch,J Expression is computed separately for moving and static faces *Neuropsychologia,31,* 173-181, 1993

8 Johansson, G. (1973) Visual perception of biological motion and a model for its analysis *Percep'n and P'physics, 14,* 201-211, 1973

9 Mace,K. & Pentland,A Lipreading by optical flow (in Japanese) *IEICE Transactions J73-D-II, 6* 796-803, 1990

10 McGurk, H. and MacDonald, J. Hearing lips and seeing voices. *Nature, 264,* 746-748., 1976

11 McLeod,P., Zihl,J., Perrett,D.I. et al. The perception of biological motion in the motion-blind patient *Visual Cognition,* in press

12 Milner,A.D., Perrett,D.I., Johnston,R.S., et al. Perception and action in visual form agnosia *Brain, 114,* 405-408, 1991

13 Regard,M., Campbell, R. & Landis, T. Tests of speechreading in neurological patients (unpub'd)

14 Rosenblum,L.D. & Saldaña,H.M. Visual primitives for audiovisual speech integration *JEP, HP&P,* 1996

15 Rosenblum,L.D., Johnson,J.A. & Saldaña,H.M. Visual kinematic information for embellishing speech in noise, submitted

16 Sergent,J Brain Imaging Studies of Cognitive Functions *TINS, 17,* 221-227 , 1994

17 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

18 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

19 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

20 Summerfield,A.Q. Use of Visual information for phonetic processing *Phonetica,36,* 314-331, 1979

21 Troscianko,T., Davidoff,J., Humphreys,G. et al. Human Colour Discrimination based on a non-parvocellular pathway *Current Biology, 6,* 1996

22 Vatikiotis-Bateson,E. & Munhall,K. In D.Stork & M.Henneke (Eds) *Speechreading by Man and machine: a NATO symposium,* Springer, Berlin , 1996

23 Vitkovitch,M. & Barber,P (1994) Effect of Videoframerate on shadowing *JSHR, 37* , 1994

24 Vitkovitch,M. & Barber, P. (1996)Visible speech as a function of image quality *Applied Cognitive Psychology,*

25 Zihl,J., von Cramon,D., Mai,N. & Schmid,C (1991) Disturbance of movement vision after bilateral posterior brain damage *Brain, 114,* 2235-52

---

[i] 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.