Dynamic visual speech perception in a patient with visual form agnosia

K. G. Munhall,[1,2,3,CA] P. Servos,[4,5] A. Santi[4] and M. A. Goodale[5,6]

Departments of [1]Psychology and [2]Otolaryngology, Queen's University, Kingston, Ontario, Canada; [3]ATR Human Information Science Laboratories, Kyoto, Japan; [4]Department of Psychology, Wilfrid Laurier University, Waterloo, Ontario; [5]CIHR Group on Action and Perception; [6]Department of Psychology, University of Western Ontario, London, Ontario, Canada

[CA,1] Corresponding Author and Address: munhallk@psyc.queensu.ca

Received 5 May 2002; accepted 5 July 2002

To examine the role of dynamic cues in visual speech perception, a patient with visual form agnosia (DF) was tested with a set of static and dynamic visual displays of three vowels. Five conditions were tested: (1) auditory only which provided only vocal pitch information, (2) dynamic visual only, (3) dynamic audiovisual with vocal pitch information, (4) dynamic audiovisual with full voice information and (5) static visual only images of postures during vowel production. DF showed normal performance in all conditions except the static visual only condition in which she scored at chance. Control subjects scored close to ceiling in this condition. The results suggest that spatiotemporal signatures for objects and events are processed separately from static form cues. *NeuroReport* 13:1793–1796 © 2002 Lippincott Williams & Wilkins.

Key words: Audiovisual speech perception; Visual form agnosia; Visual motion; Face perception

INTRODUCTION

Recent evidence indicates that motion can play different roles in object perception. Motion can be used to define the contours of an object [1] but specific patterns of motion or spatiotemporal signatures [2] also provide information for object and event recognition independent of form and contour. In the latter case the motion itself is a characteristic of the object or event. Animate motion is a classic example of this dynamic object perception. Humans and other species can distinguish animate motion from other motions when form information is degraded or absent [3] and can even identify a wide range of activities that are being performed using only the visual kinematics. The relationship between motion-defined object recognition and recognition that relies on static form attributes is not well understood. In this paper we explore the functional neuroanatomy of static and dynamic perception by testing a patient with visual form agnosia on visual speech stimuli.

Speech perception is both a visual and an auditory process and it is well known that the intelligibility of auditory speech perception is enhanced when a listener can view the talker's face. This enhancement occurs whether the auditory degradation is due to background noise, sensorineural hearing loss or communication in a second language. In some cases visible articulation can even change the perception of perfectly audible speech [4] or be used as a substitute for auditory speech perception [5].

Both static posture cues and the movements of the face and head during articulation are potential sources of this visual linguistic information. Static configurations of the face can be used to distinguish a range of vowels and consonants [6] and subjects can reliably label many speech sounds based on static photographs of a face producing the sound [7]. On the other hand, a variety of data suggest that the dynamic information from the face contributes independently to a range of perceptual judgments about speech, emotion and identity. The movement of the head is a major visual cue for the prosody of sentences and head motion is known to correlate strongly with the acoustic pitch of the voice [8,9]. The movement of the soft tissue of the face provides the primary visual information for individual speech sounds and this non-rigid motion is distributed across the face [8]. The motion of different parts of the facial surface contributes independently to prediction of the acoustic spectrum and RMS amplitude of the acoustic speech signal [9]; perceptually this distributed facial information influences auditory speech intelligibility.

In the absence of any static configuration information, point-light displays can be used to enhance speech perception in noise [10] and to produce the McGurk effect [11]. Subjects can also use point-light kinematics to classify emotional expressions [12] and to some extent determine the gender and identity of an individual [13]. The perception of the identity of degraded images of famous faces or personal acquaintances can be enhanced significantly with motion [14] and subjects can use motion cues to distinguish the gender and identity of animations. In a series of experiments, Hill and Johnston [15] animated a single rendering of a face and head with the recorded motions of different actors. Thus the facial features remained the same but the motion characteristics varied for the different actors. Subjects used both the rigid motion of the head and the non-rigid motion of the face to determine the gender and identity of the actors.

The neurospsychology of visual speech perception supports the distinction between dynamic and static speech processing. Humphreys *et al.* [16] describe patient HJA's impairment at recognizing familiar faces and judging gender and emotional expression from static photographs. However HJA could judge expression and gender from point-light displays. Campbell [17] reports that HJA could also correctly identify moving video images of speech but could not identify static photographs of the speech sounds being produced. In contrast, patient LM [18] showed preserved static speech recognition but gross impairment in tasks in which dynamic speech stimuli are presented.

In this paper we add to this small body of literature by presenting data from a patient with selective damage to the ventral stream of visual processing [19]. While it is well documented that patient DF has profound visual form processing deficits with preserved use of vision for the control of action, less is known about her ability to use dynamic information to make categorical judgments such as required in a speech perception task. We tested DF and control subjects on simple speech perception tasks that require visual speech processing including tasks that require dynamic *vs* static image processing.

MATERIALS AND METHODS

Subjects: Patient DF is a 46-year-old woman who at age 34 suffered irreversible brain damage as a consequence of carbon monoxide poisoning. MRI that was carried out ~1 year following the accident indicated damage ventrally in the parasagittal occipitoparietal region (primarily Brodmann's areas (BA) 18 and 19) but with apparent sparing of areas 17, 20, 21 and 37. For a detailed description of DF's brain damage including MRI images see [20]. Subsequent neuropsychological and psychophysical testing revealed the presence of a profound visual form agnosia. DF was impaired in the perception of shape and orientation regardless of which stimulus parameters were used to define the contours (intensity, colour, texture, stereopsis, motion, proximity, continuity, or similarity). DF showed no ability to visually recognize familiar faces though she can use context and voice quality to identify individuals [20]. Psychophysical testing revealed that her visual form agnosia could not be reduced to a simple sensory deficit.

Control subjects: A 40-year-old right-handed female (EW) served as an age-matched control and a group of 12 undergraduates at Queen's University served as additional controls. All controls were native speakers of English with no known speech, language or neurological problems and had normal or corrected-to-normal vision.

Stimuli: The audiovisual stimuli were taken from the Johns Hopkins University audiovisual laserdisc [21]. Two American English speakers produced the point vowels /i/, /u/, and /a/. For the audio only condition and one of the audiovisual conditions the audio signal was an electroglottograph (EGG) recording of the talker's voice. The EGG transduces vocal fold contact area [22] and thus is an indication of when the talker is producing voicing and also an indication of the pitch or fundamental frequency of their voice (F0). When this signal is played through a speaker it sounds like a muffled vocal pitch but it contains minimal spectral information about which vowel is being spoken. The other audio signal was a sound recording of the talkers' voices which contained the full vowel spectrum. Two video signals were used. The first was the dynamic video of the talkers saying the vowels. The other video signal was a single frame taken from the dynamic video. The frame chosen was the extreme position of the lips during the vowel articulation. The single frame was displayed for a duration equal to the duration of the dynamic vowel video stimulus.

Procedure: DF and the aged-matched control EW were tested in a quiet room with a video recording of the stimuli. The undergraduate controls were tested in a double-walled sound isolation booth (IAC Model 1204) using a computer-controlled laserdisc display system. Five conditions were tested in the following order: (1) auditory only with the audio signal from the EGG; (2) audiovisual with the audio signal from the EGG; (3) video only with dynamic video; (4) audiovisual with the audio signal from an audio recording; (5) video only with a static video frame. Subjects were tested on 120 trials (5 conditions × 3 vowels × 2 talkers × 4 repetitions). The stimuli were randomized within a condition and subjects chose one of three vowel responses for each trial.

RESULTS

The percentage of correctly identified vowels for each condition is shown in Figure 1. As expected, performance for DF and the controls was approximately at chance for the auditory only pitch condition. For the three conditions with dynamic facial stimuli, DF and the controls approached or reached ceiling performance. While DF scored slightly worse than the average control scores on all of these four conditions, her accuracy level was higher than or equal to the lowest control score in each condition (25%, 75%, 83%, 100% for auditory only pitch, audiovisual pitch, visual only, and audiovisual voice respectively). However, DF's results with the static vowels demonstrated a marked deficit. Her score in this condition was at chance while the controls approached ceiling performance levels. In contrast to DF's score of 29.2%, the lowest control performance in the static vowels condition was 87.5%.

DISCUSSION

The results clearly indicate that DF has preserved dynamic speech perception in the face of impaired static form processing. Her performance on all speech perception tasks that required judgments about moving faces was quite similar to controls. However, her performance on the static vowel identification task was markedly different. Like patient HJA [16,17], DF showed no evidence of static speech perception ability and no ability to generalize from the

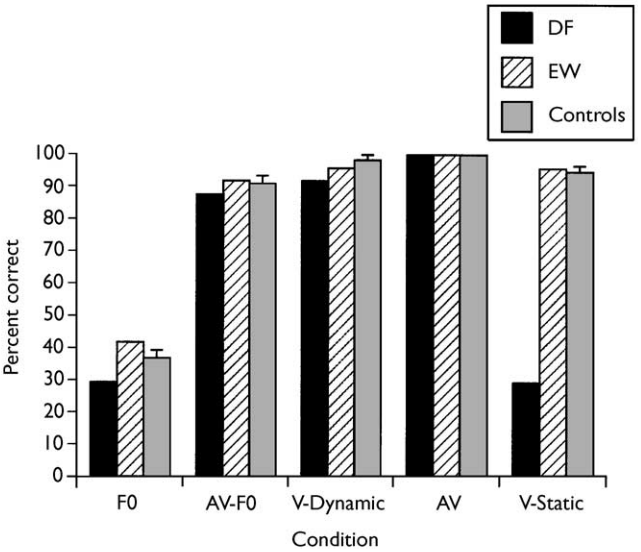

Fig. 1. The figure shows the percentage of correctly identified vowels for the five conditions (auditory only with the auditory signal conveying vocal pitch (EGG), audiovisual with the auditory signal conveying vocal pitch (EGG) with a moving face, visual only, audiovisual with the auditory signal being the recorded voice signal, and visual only with a static picture of a face producing a vowel). The results for patient DF, and aged-matched control (EW), and undergraduate controls are presented. Error bars indicate s.e.m.

dynamic facial movements to postures associated with the facial movement sequence.

Recent neuroimaging data are consistent with DF's results. Calvert *et al.* [23] compared fMRI activation patterns in subjects presented with static and dynamic visual speech stimuli. While performance on their phoneme monitoring task was approximately equal in the two visual speech conditions, more extensive activation was observed for the moving stimuli. Dynamic stimuli preferentially activated (as compared to static speech photographs) areas in the inferior frontal gyrus, superior temporal gyrus and the superior temporal sulcus. Interestingly, one of the areas that showed preferential activation for static speech photographs (as compared to dynamic speech stimuli) was the lingual gyrus (BA 18), an area that overlaps with DF's lesions. DF's lesions are located primarily in the lateral prestriate cortex including large parts of V2, V3, and V4 [19,20]. HJA has similar bilateral lesions in the occipital cortex [16] with damage to V4 being greater than V2 and V3 [18]. DF's and HJA's damage to ventral 'object' regions is in contrast to patient LM's lesions which are almost exclusively confined to V5 [18]. These neurological data suggest that motion and static form information about speech are processed differently and provide different cues about sound categories.

Studies of emotional expression and identity also indicate that facial motion is not redundant with static facial configuration. Kamachi *et al.* [24], using dynamically morphed facial expressions, showed that the velocity of facial motion influenced the accuracy and judgments of the intensity of emotional expression. Neither the overall duration of the stimulus nor the static endpoint expression could account for subject ratings. Lander and Bruce [14] have demonstrated that motion enhances the identification of familiar faces across a range of conditions in which the spatial cues are severely degraded (e.g. negative images, inversion, and various spatial frequency manipulations such as pixelation and blurring). Given the robustness of subjects' performance in these tasks, dynamic information must be an integral part of the perception of speech, emotion and identity. Like the role of color and other surface properties in object perception [25], motion patterns may be an indexing feature for particular objects and events.

A range of different types of motion and demands on the motion processing systems are a part of everyday perception. As Stone [2] suggests, a continuum of spatiotemporal stimuli exists ranging from non-rigid deformation to articulated motion of rigid parts (e.g. walking figures) to rigid objects that move as a whole. Whether this continuum of motion is processed in the same manner and how static object features are integrated are unknown. There are indications, however, that the nervous system may process some classes of motion with specialized neural substrates [26]. Understanding the perception of this motion and its integration with texture and surface processing will rely on converging evidence from clinical populations and functional imaging.

CONCLUSION

The results of the present study indicate that dynamic visual speech perception is distinct from the perception of static facial form. DF's deficits suggest that the network involved in natural audiovisual speech perception is separate from the ventral stream of visual processing.

REFERENCES

1. Wallach H and O'Connell D. *J Exp Psychol* **45**, 205–217 (1953).
2. Stone JV. *Vis Res* **38**, 957–951 (1998).
3. Johansson G. *Percept Psychophys* **14**, 201–211 (1973).
4. McGurk H and MacDonald J. *Nature* **264**, 746–748 (1976).
5. Bernstein LE, Demorest ME and Tucker PE. *Percept Psychophys* **62**, 233–252 (2000).
6. Fromkin VA. *Lang Speech* **7**, 215–225 (1964).
7. Campbell R, Landis T and Regard M. *Brain* **109**, 509–521 (1986).
8. Vatikiotis−Bateson E, Munhall KG, Hirayama M *et al.* Physiology−based synthesis of audiovisual speech. In: *Proceedings of 4th Speech Production Seminar: Models and Data*. Autrans, France. Grenoble: Institut de la Communication Parlée; 1996, pp. 241–244.
9. Yehia HC, Rubin PE and Vatikiotis-Bateson E. *Speech Commun* **26**, 23–44 (1998).
10. Rosenblum LD, Johnson JA and Saldaña HM. *J Speech Hear Res* **39**, 1159–1170 (1996).
11. Rosenblum LD and Saldaña HM. *J Exp Psychol Hum Percept Perf* **22**, 318–331 (1996).
12. Bassili JN. *J Exp Psychol Hum Percept Perf* **4**, 373–379 (1978).
13. Bruce V and Valentine T. When a nod's as good as a wink. The role of dynamic information in facial recognition. In: Gruneberg MM and Morris E, eds. *Practical Aspects of Memory: Current Research and Issues*. New York: Lawrence Erlbaum Associates; 1988, pp. 169–174.
14. Lander K and Bruce V. *Ecol Psychol* **12**, 259–272 (2000).
15. Hill H and Johnston A. *Curr Biol* **11**, 880–885 (2001).
16. Humphreys GW, Donnelly N and Riddoch MJ. *Neuropsychologia* **31**, 173–181 (1993).
17. Campbell R. *Phil Trans R Soc Lond B* **335**, 39–45 (1992).
18. Campbell R, Zihl J, Massaro D *et al. Brain* **120**, 1793–1803 (1997).
19. Milner AD and Goodale MA. *The Visual Brain in Action*. Oxford: Oxford University Press; 1995.
20. Milner AD, Perrett DI, Johnston RS *et al. Brain* **114**, 405–428 (1991).

21. Bernstein LE and Eberhardt S. Audiovisual Laserdisc. Department of Electrical and Computer Engineering, Johns Hopkins University (1986).
22. Childers DG and Krishnamurthy AK. *Crit Rev Biomed Eng* **12**, 131–161 (1985).
23. Calvert GA, Brammer MJ and Campbell R. *Cortical substrates of seeing speech: Still and moving faces*. Paper presented at Human Brain Mapping 2001, Brighton, UK.
24. Kamachi M, Bruce V, Mukaida S *et al. Perception* **30**, 875–887 (2001).
25. Humphrey GK, Goodale MA, Jakobson LS and Sevos P. *Perception* **23**, 1457–1481 (1994).
26. Vaina LM, Solomon J, Chowdhury S *et al. Proc Natl Acad Sci USA* **98**, 11656–11661 (2001).

Acknowledgements: This work was funded by grants from NIDCD, CIHR, NSERC, CRL Keihanna Info–Communications Research Laboratories and by the Telecommunications Advancement Organization of Japan.