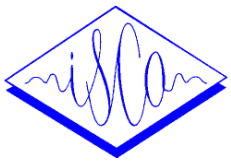

ESCA Workshop on
Audio-Visual Speech Processing
Rhodes, Greece
September 26-27, 2997

IMPAIRMENT OF VISUAL SPEECH INTEGRATION IN PROSOPAGNOSIA

Beatrice de Gelder*, Nancy Etcoff# and Jean Vroomen*
*Tilburg University and #Harvard Medical School

Beatrice de Gelder, Dept of Psychology, Tilburg University
PObox 90153, 5000 LE Tilburg, The Netherlands
e-mail: b.degelder@kub.nl phone:+33-(0)13-4662167

ABSTRACT

Our study of a prosopagnosic patient LH shows a strong association between severe face processing deficits and loss of speechreading skills. With simple dynamic stimuli some speechreading ability seems preserved but it is insufficient to affect the processing of auditory input and to generate audiovisual blends or to provoke cross-modal bias.

1. INTRODUCTION

Personal identity, age, gender, emotion and speech can all be read from the face. But the face is not the exclusive bearer of all that information. For example, the voice can often be just as informative about gender or emotion of the speaker. Reports of face processing impairments (prosopagnosia) have raised the question whether all kinds of information carried by the face would be impaired in prosopagnosic patients (see Damasio, Tranel & Damasio, 1990 for an overview). An equally legitimate question is whether in the presence of an impairment in visual speech processing, the auditory input channel continues to function fully and operate independently with bi-modal input. For example, if speechreading skills seem lost as a consequence of prosopagnosia, does the patient process auditory speech normally when presented with a combined voice and face stimulus? We explored some of these issues with LH, a well known prosopagnosic patient. (Etcoff, Freeman & Cave, 1991).

Speechreading appears to be a good candidate for an ability that could be preserved in patients having lost access to other aspects of facial semantics. The co-occurrence of prosopagnosia and right hemisphere lesions together with the well known implementation of language ability in the left hemisphere provides a neurobiological framework in which dissociations between speechreading and face processing might be expected. The first report of just such a dissociation was offered by Campbell, Landis and Regard(1986). Patient Mrs D. was densely agnosic with profound prosopagnosia, yet could sort pictures of faces according to speechsound and was sensitive to the effects of seeing the speaker in reporting heard speech (McGurk effects). She could speechread silent spoken numbers as well as discriminate lipspoken vowels and consonants. By contrast, patient Mrs T. was unable to perform such tasks, although she had no difficulty recognizing faces or facial expressions or other visual objects but was alexic. Mrs T's lesion was unilateral and affected the left hemisphere, Mrs D's only affected the right. More recently a study of HJA (Campbell, 1992), a patient with bilateral lesions of occipito-temporal areas, showed prosopagnosia and could not classify photographs of speaking faces. He was however completely normal with dynamic stimuli and presented normal audiovisual integration.

The importance of visual movement pathways to speechreading is illustrated by patient LM (Campbell, 1996). LM's lesion affects only the cortical visual movement areas, including area V5, and sparing areas V1-V4 which are all damaged in HJA. LM can only classify still photographs and does not show McGurk effects. This pattern suggests a dissociation between static and dynamic inputs to speechreading which would imply that at least in some basic sense, both the perception of static forms as well as the perception of movement patterns can access speech representations.

An extensive study of speechreading ability in a visual agnosic patient was presented by de Gelder, Vroomen and Bachoud-Lévi (1997). Their patient BC was impaired in all aspects of face processing. Auditory language processing was normal but speechreading from still faces was entirely lost. There was no indication of a movement perception disorder in this patient and indeed, with dynamic displays of talking faces she yielded a better speechreading score than with still faces but she was still performing below normal. Often the movement of the lips was perceived in a systematic fashion but not related to the correct phoneme. Since some speechreading ability with dynamic displays was preserved we expected a reasonably normal performance on tasks that test the integration of auditory and visual speech like those examining audio-visual bias and audio-visual integration. Instead, we found no evidence for this partly preserved dynamic speechreading ability in bi-modal situations. This suggests that BC has a route to speechreading but its output does not merge with auditory speech, either because they are

too weak or too coarse, or because her partly preserved speechreading is qualitatively different from normal. The finding of a spared ability for speechreading from motion observed in HJA does clearly not generalize to BC, although there are many similarities between the two patients in the site of the lesions as well as in object and face recognition impairments and there is no movement perception disorder per se in BC.

Below we present speechreading abilities of a prosopagnosic patient LH. His face processing impairments have extensively been reported on, most recently by Etcoff, Freeman and (1991), Farah et al.(1995) and de Gelder and Etcoff (1997). LH is profoundly prosopagnosic (with some mild visual object agnosia) with intact abilities in the domain of language. Besides the bi-modal experiments reported below, we tested LH with a task of speechreading from still faces, with a single digit speechreading task and with an audiovisual memory task. In what follows we report specifically the bi-modal tasks. The results of the latter will be discussed against the background of other speechreading tasks with which LH was examined so far.

2. METHOD AND RESULTS

Task 1: Auditory processing, speechreading and audio-visual conflict. We used a video recording of a female speaker pronouncing a series of VCV sequences(de Gelder, Vroomen & van der Heide, 1991; de Gelder, Vroomen & Bachoud-Lévi, 1997). Each sequence consisted of one of the four plosive stops /p, b, t, d/ or a nasal /m, n/ in between the vowel /a/ (e.g., /aba/ or /ana/). There were three presentation conditions: an audio-visual, an auditory-only, and a visual-only presentation. In the audio-visual presentation, dubbing operations were performed on the recordings so as to produce a new video-film comprising six different auditory-visual combinations: auditory /p, b, t, d, m, n/ were combined with visual /t, d, p, b, n, m/, respectively. The visual place of articulation feature thus never matched the auditory place feature. Appropriate dubbing ensured that there was auditory-visual coincidence of the release of the consonant in each utterance. In addition, unimodal presentation conditions were produced. For the audio-only condition, the original auditory signal was dubbed on a blank screen. For the visual-only condition, the auditory channel was deleted from the recording, so the subject had to rely entirely on speechreading. Each presentation condition comprised three replications of the six possible stimuli. LH was instructed to watch the speaker and repeat what she said.

Results: In the audio-visual conflict condition, there were only two fusions out of 18 trials (11%) while normal performance is about 50% (see de Gelder, Vroomen & van der Heide, 1991). In all other trials he reported the audio-part of the audio-visual stimulus. In the auditory-only trials he always reported the correct phoneme. For the visual-only trials, two response categories were made, based on two broad viseme classes: lingual (d, t, or n) or bi-labial (b, p, m). Performance was in this case only 22% correct. On the nine bi-labial trials, he reported two times a bi-labial, the other times a lingual. On the nine lingual trials, he reported two times a lingual, and seven times a bilabial.

LH is thus normal at processing the auditory input presented on its own in the absence of a face. But he performs poorly when having to report what is said by a face in the absence of any auditory input. Therefore it is not surprising that he does not show fusions or blends and only tends to report the audio part of a bimodal stimulus when there is a conflict between the information in audition and vision. Such a result looks straightforward and implies that his prosopagnosia strongly affects his speechreading ability and that notwithstanding normal movement processing LH cannot process visual speech, however presented..

Task 2: Synthetic face and voice The next task focuses on audio-visual bias and should offer an opportunity for a more fine grained appraisal of separate as well as combined processing in the two speech input modalities. The task is a variant of the well known categorical perception paradigm, and requires the use of synthetic speech as well as a synthetic face. Normal subjects presented with synthetic speech or a synthetic talking face show more variability in their performance. But no systematic comparison is available of the performance on natural versus synthetic stimuli and it is not clear how the use of synthetic stimuli like faces or voices might affect performance of patients with focal brain damage and selective functional deficits. We will return to this caveat in the discussion. Like the previous task the materials consist of bimodal as well as unimodal trials. But an important difference is that unlike in the previous case, the unimodal auditory trials always consist of a speech stimulus combined with a still face. This allows us to appreciate among other things whether the auditory speech channel is autonomous and still robust enough in the presence of a face.

The task consisted of a tape showing an artificially created synthesized face (Massaro & Cohen, 1990). The synthetic face is controlled by 11 display parameters which determine jaw rotation, lip protrusion, upper lip raise, etc. By varying these parameters, a dynamic face is created that articulates 'ba', 'da' or any intermediate position between these two syllables. In the test, five levels

of audio speech varying between 'ba' and 'da' were crossed with 5 levels of visible speech varying between 'ba' and 'da'. These 25 stimuli comprise the audiovisual condition. The auditory and visual stimuli were also presented alone, so that there was a total of 25 + 5 + 5 = 35 independent stimulus events. The whole test consisted of 6 sets of these 35 trials in which the order of items was randomized.

Results Task 2. The performance of LH was compared with that of four control subjects belonging to a similar age range. All participants were instructed to listen and to watch the video and to identify each token as 'ba', 'da', 'bda', 'dba', 'va', 'tha', 'ga' or 'other'. There were thus 8 response possibilities x 35 trial types = 280 categories. In order to decrease this number, we scored the number of 'ba'- and 'bda'-responses as one category, and 'da'- and 'tha'-responses as another category, because these categories are visually very similar and they accounted for 80% of BC's judgements. We then computed 4 different performance measures: the visual and auditory influence in the bimodal condition, and the percentage correct in visual-only and auditory-only trials. For the visual influence in bimodal trials, a visual 'ba' (i.e., the first two levels of the visual 'ba-da' continuum) should, - compared with visual 'da' - increase the number of 'ba'- and 'bda'-responses, and a visual 'da' (i.e., the final two levels of the visual 'ba-da' continuum) should - compared with visual 'ba' - increase the number of 'da'- or 'tha'-responses. The bigger these differences, the bigger the visual influence in audio-visual trials. The same logic was applied to the auditory influence in bimodal trials. An auditory 'ba' (i.e. the first two stimuli of the auditory 'ba-da' continuum) should, compared to auditory 'da' (i.e. the final two stimuli of the auditory 'ba-da' continuum), increase the number of ba- and 'bda'-responses, and an auditory 'da' should, compared to auditory 'ba' increase the number of 'da'-responses. The difference should give an indication of the auditory influence in audio-visual trials. For the visual-only trials, we computed the number of correct identifications. That is, the number of 'ba'- or 'bda'-responses when the first two stimuli of the visual 'ba-da' continuum were presented, and the number of 'da'- or 'tha'-responses when the final two stimuli of the visual 'ba-da' continuum were presented. For the auditory-only trials, we computed the number of 'ba'- and 'bda'-responses when auditory 'ba' was presented, and the number of 'da'- and 'tha'-responses when auditory 'da' was presented.

LH's had a negative visual influence in bimodal trials -14% vs. 26% (range 8%-43%) for control subjects. His auditory influence in bimodal trials was however normal: 37% for LH vs. 26% (range 8%-43%) for control subjects. In the visual-only trials LH performed surprisingly well: 58% correct vs. 55% (range 46%-67%) for the controls. His speechreading performance with artificial visual stimuli is thus superior to that observed in the previous task using a natural speaker. But in contrast to what was observed previously, LH performed poorly on the auditory-only trials(17% correct for LH vs. 64% correct for control). As we noted, the difference between auditory trials of this task and those of the previous one is that one contains natural speech, the other synthetic, and that a face was present on the screen in task 2, but not in task 1.

The result on bi-modal trials converges with that obtained with the previous task. Like in the first task, we find that visual influence in bimodal trials is non-existent.

The two tasks reported here have predominantly addressed issues of bi-modal integration. For a better understanding of the results just described information from other speech reading tasks is also relevant. In our second task with the synthetic face LH showed some preserved speechreading skill. This performance contrasts with his inability to process any information about speech provided by the form of the lips presented by still photographs. It thus seems to be the case for LH (like for BC) that some speechreading information is supported by dynamic displays but that it is insufficient. Another task done with LH was recognition of single silently spoken digits (1 to 9). His performance was above chance but much below that of normal controls. This single digit recognition task served as a control for the results obtained on a subsequent task. LH was tested with a serial recall task presented on a video tape and consisting of a number of lists each 8 digits long. Three conditions were presented, audio only, visual only and audiovisual. LH was near perfect on the audio and the audiovisual list but his performance was very poor on the visual-only lists on which controls where overall still above 80% correct. If we take this result together with the data obtained on the single digit visual recognition, it is clear that the poor performance on the serial recall task is not due to memory problems, but most likely to the fact that the visual speech representations are too impoverished and weak to sustain phonological memory storage.

3. DISCUSSION

The case of LH clearly underscores that face processing and speechreading are both seriously affected in prosopagnosics like BC and that the close link between auditory and visual input for speech in normals does not safeguard

speechreading ability in case of a generalized face recognition disorder. The study of LH raises several issues that need to be pursued and several aspects of the speechreading impairment in prosopagnosics that require attention. An important issue concerns the exact source of the speechreading impairment of prosopagnosics. Cases like LH and BC raise the issue of the locus of the speechreading impairment and the question where in the processing of the visual speech information, face recognition processes interfere with visual language processing. The synthetic face used in the second task presents a stimulus where all detail is left out. One might argue that this allows an easier perceptual segregation of the mouth from the rest of the face and improves perception of the dynamics of lipmovements. But when this possibility was tested here (as well as previously with BC) it turned out that without the full facial context present, speechreading was even more difficult. As to the difference between natural and synthetic faces, no data are available. Such an improvement from a natural to a synthetic speechreading task was not observed in BC. In any event, this good speechreading performance makes the total absence of visual bias all the more surprising.

Like in the case of BC, it seems that some preserved speechreading ability does not predict that audiovisual conflict will occur. Like in the case of BC this might be a consequence of too weak or too coarse visual speech representations. Alternatively, the partly preserved speechreading ability might be achieved by a non- specific route, for example by exclusive focus on the lipmovements without contribution from linguistic representations.

Finally, we note the intriguing fact that auditory recognition is severely handicapped by the presence of a still face. The same phenomenon was observed in BC. Further research is needed to decide whether this might be a consequence of an expectancy bias, of a cross-modal bias, or more interestingly, an interference from preserved face perception (but not recognition) abilities. Both LH (de Gelder and Etcoff, 1997) and BC (de Gelder, Bachoud-Lévi and Degos, 1996) have lost face recognition but show evidence for spared structural encoding of faces and continue to perceive faces as faces. The interference from preserved structural encoding of faces observed previously in our face matching and recognition tasks might thus extend to the processing of visual speech. Preserved structural face perception might interfere with the recognition of speech sounds when these are presented in the context of a still face.

4. REFERENCES

Campbell, R., Landis, T., & Regard, M. (1986). Face recognition and lipreading: A neurological dissociation. Brain, 109, 509-521.

Campbell, R. (1992) The neuropsychology of lipreading. Phil. Trans. Roy. Soc. London, B. 335, 39-45.

Campbell, R. (1996) Seeing speech in space and time. Proceedings of the 4th International Conference on Spoken Language Processing. Philadelphia, October.

Damasio, A. R., Tranel, D., & Damasio, H. (1990). Face agnosia and the neural substrates of memory. Annual Review of Neuroscience, 13, 89-109.

Etcoff, N. L., Freeman, R., & Cave, K. L. (1991) Can we lose memories of faces? Content. awareness in a prosopagnosic.

Farah, M., Levinson, K., & Klein, K. (1995a). Face perception and within-category discrimination in prosopagnosia. Neuropsychologia, 33(6), 661-674.

Gelder, B. de, Vroomen, J., & van der Heide, L. (1991). Face recognition and lipreading in autism. European Journal of Cognitive Psychology, 3, 69-86.

Gelder, B. de, Bachoud-Lévi, A.-C. & Degos, J.D. (1996) Objects and faces: An inverted face and object effect combined in a case of prosopagnosia. Brain and Cognition, vol. 32, p. 269.

Gelder, B. de & Vroomen, J. (1996). Auditory illusions as evidence for a role of the syllable in adult developmental dyslexics. Brain and Language, 52, 373-385.

Gelder, B. de & Etcoff, N. (1997) Component-based strategies do not compensate for loss of face and object recognition. Meeting of the Psychonomic Society, Inc. Philadelphia, November 20-23.

Massaro, D.W. & Cohen, M.M. (1990). Perception of synthesized audible and visible speech. Psychological Science, 1, 1-9.