Perception of co-speech gestures in aphasic patients: A visual exploration study during the observation of dyadic conversations

Basil C. Preisig [a], *Noëmi Eggenberger* [a], *Giuseppe Zito* [b], *Tim Vanbellingen* [a,c], *Rahel Schumacher* [a], *Simone Hopfner* [a], *Thomas Nyffeler* [a,c], *Klemens Gutbrod* [d], *Jean-Marie Annoni* [e], *Stephan Bohlhalter* [a,c] and *René M. Müri* [a,d,f,g,*]

[a] *Perception and Eye Movement Laboratory, Departments of Neurology and Clinical Research, Inselspital, University Hospital Bern, and University of Bern, Switzerland*

[b] *ARTORG Center for Biomedical Engineering Research, University of Bern, Switzerland*

[c] *Neurology and Neurorehabilitation Center, Department of Internal Medicine, Luzerner Kantonsspital, Switzerland*

[d] *Division of Cognitive and Restorative Neurology, Department of Neurology, Inselspital, Bern University Hospital, and University of Bern, Switzerland*

[e] *Neurology Unit, Laboratory for Cognitive and Neurological Sciences, Department of Medicine, Faculty of Science, University of Fribourg, Switzerland*

[f] *Gerontechnology and Rehabilitation Group, University of Bern, Bern, Switzerland*

[g] *Center for Cognition, Learning and Memory, University of Bern, Bern, Switzerland*

Background: Co-speech gestures are part of nonverbal communication during conversations. They either support the verbal message or provide the interlocutor with additional information. Furthermore, they prompt as nonverbal cues the cooperative process of turn taking. In the present study, we investigated the influence of co-speech gestures on the perception of dyadic dialogue in aphasic patients. In particular, we analysed the impact of co-speech gestures on gaze direction (towards speaker or listener) and fixation of body parts. We hypothesized that aphasic patients, who are restricted in verbal comprehension, adapt their visual exploration strategies.

Methods: Sixteen aphasic patients and 23 healthy control subjects participated in the study. Visual exploration behaviour was measured by means of a contact-free infrared eye-tracker while subjects were watching videos depicting spontaneous dialogues between two individuals. Cumulative fixation duration and mean fixation duration were calculated for the factors co-speech gesture (present and absent), gaze direction (to the speaker or to the listener), and region of interest (ROI), including hands, face, and body.

[*] *Corresponding author*. Perception and Eye Movement Laboratory, Departments of Neurology and Clinical Research, Inselspital, University Hospital Bern, Freiburgstrasse 10, 3010 Bern, Switzerland.

E-mail address: Rene.mueri@insel.ch (R.M. Müri).

Results: Both aphasic patients and healthy controls mainly fixated the speaker's face. We found a significant co-speech gesture × ROI interaction, indicating that the presence of a co-speech gesture encouraged subjects to look at the speaker. Further, there was a significant gaze direction × ROI × group interaction revealing that aphasic patients showed reduced cumulative fixation duration on the speaker's face compared to healthy controls.

Conclusion: Co-speech gestures guide the observer's attention towards the speaker, the source of semantic input. It is discussed whether an underlying semantic processing deficit or a deficit to integrate audio-visual information may cause aphasic patients to explore less the speaker's face.

1. Introduction

Co-speech gestures can be defined as hand movements that accompany spontaneous speech and they are thought to have a nonverbal communicative function (Kendon, 2004). Nonverbal behaviour in humans is most often idiosyncratic, meaning that in contrast to verbal language no common lexicon for gestural expression exists. Therefore, a wealth of classification systems for co-speech gestures has emerged over time (Lott, 1999). Co-speech gestures can be redundant (e.g., pointing while naming an object), supplementary (e.g., shrug to express one's uncertainty), or even compensatory to direct speech (e.g., ok sign). In addition, they were also found to facilitate lexical retrieval (Krauss & Hadar, 1999) and to complement speech prosody (Krahmer & Swerts, 2007).

Aphasia is an acquired language disorder that occurs as a consequence of brain damage to the language dominant hemisphere. It is a disorder with supra-modal aspects that commonly affects both production and comprehension of spoken and written language (Damasio, 1992). The disorder may be explained from a language-based or from a cognitive processing view. The language-based, clinically oriented approach assumes that neural damage directly affects specific language functions causing linguistic deficits on the phonological, syntactical, and semantic level of language processing. The cognitive view suggests that aphasic symptoms are caused by impaired cognitive processes which support language construction. These cognitive processes can be understood as a specialized attentional or memory system which is vulnerable to competing input from other processing domains (Hula & McNeil, 2008).

Previous work in aphasic patients focused on gesture production and presented conflicting evidence. Some studies suggest that patients communicate better if they use gestures (Behrmann & Penn, 1984; Herrmann, Reichle, Lucius-Hoene, Wallesch, & Johannsen-Horbach, 1988; Lanyon & Rose, 2009; Rousseaux, Daveluy, & Kozlowski, 2010); others claim that the ability to use gestures and to speak breaks down in parallel in aphasia (Cicone, Wapner, Foldi, Zurif, & Gardner, 1979; Duffy, Duffy, & Pearson, 1975; Glosser, Wiener, & Kaplan, 1986). There are different explanations for the inconsistency of findings: Rimé and Schiaratura (1991) suggested that it is difficult to compare the results of different studies, because the authors provided their own solution to handle gesture classification. Furthermore, co-occurrence of apraxia, an impairment of the ability to perform skilled, purposive limb movements (Ochipa & Gonzalez Rothi, 2000), has often been neglected in studies on gesture production.

The analysis of visual exploration provides insights for the understanding of gesture processing. Moreover, the recording of eye movements has proven to be a valid and reliable technique to assess visual exploration behaviour (Henderson & Hollingworth, 1999). Previous studies analysed healthy subjects' visual exploration of co-speech gestures while observing an actor who was retelling cartoon stories. These studies found that gestures attract only a minor portion of attention (2–7%), while the speaker's face is much more fixated (90–95%) (Beattie, Webster, & Ross, 2010; Gullberg & Holmqvist, 1999, 2006; Gullberg & Kita, 2009; Nobe, Hayamizu, Hasegawa, & Takahashi, 2000). To the best of our knowledge the visual exploration behaviour of co-speech gestures has not been studied in aphasic patients.

In the present study, we were interested in the visual exploration of a dyadic dialogue condition. Dyadic dialogue can be defined as two people who are engaged in a conversation. In contrast to monologue, which can stand for itself, dialogue depends on the collaboration between the interlocutors (Clark & Wilkes-Gibbs, 1986) and requires processes such as the organization of turn taking (Sacks, Schegloff, & Jefferson, 1974). In this study, we presented spontaneous dyadic dialogues on video while visual exploration behaviour of aphasic patients and healthy controls was assessed by means of an infrared eye-tracking device. Previous research in multiparty conversations suggests that people most likely look at the person who is speaking or whom they are speaking to (Vertegaal, Slagter, van der Veer, & Nijholt, 2000). In addition, Hirvenkari et al. (2013) reported that after a turn transition the gaze is directed towards the speaking person. Therefore it could be assumed that non-involved observers are also inclined to look at the speaker, while following the dyadic conversation. Moreover, we were interested whether co-speech gestures have an additional influence on gaze direction. Thus, our first hypothesis is that co-speech gestures modulate gaze direction of the observer towards the speaking actor in the video. Since auditory speech perception can be affected in aphasia (Hickok & Poeppel, 2000) patients may rely more on other communication channels which results in a modified visual exploration pattern of face and hand region. Thus the second hypothesis is that aphasic patients show different visual exploration patterns of the face and the hand region compared to healthy controls, allowing them either to compensate for language impairment or to avoid interference between the visual and the auditory speech signal. We propose three different visual exploration strategies: It is known from the literature (Arnal, Morillon, Kell, & Giraud, 2009; van Wassenhove, Grant, & Poeppel, 2005) that viewing articulatory movements of the speaker's face facilitates auditory speech perception. Therefore, aphasic patients may fixate the speaker's face more – thus focusing on the visual speech signal – in order to compensate auditory comprehension deficits. A second suggestion is that aphasic patients may compensate auditory comprehension deficits with additional nonverbal information deriving from the actors' co-speech gestures and therefore may fixate more the speaker's co-speech gestures, i.e., the hands. There is evidence that the presence of co-speech gestures improves information encoding and memory consolidation (Cohen & Otterbein, 1992; Cook, Duffy, & Fenn, 2013; Feyereisen, 2006; Records, 1994). Finally, the third suggestion is based on the cognitive processing view: Aphasic patients allocate their limited attentional resources (Kahneman, 1973) on the auditory input and avoid competing input from visual speech perception. Furthermore, there are indications for an audio-visual integration deficit in aphasic patients (Schmid & Ziegler, 2006; Youse, Cienkowski, & Coelho, 2004). Hence, it is suggested that aphasic patients fixate the speaker's face less.

2. Material & methods

2.1. Subjects

Sixteen patients with left hemispheric cerebrovascular insult (aged between 34 and 74, $M = 52.6$, $SD = 13.3$, 5 females, 1 left-handed) and 23 healthy controls (aged between 23 and 73, $M = 50.3$, $SD = 16.4$, 8 females, 1 left-handed, 1 ambidexter) participated in the study. There was no statistically significant difference between the groups with respect to age [$t(37) = .459$, $p = .649$, 2-tailed] and gender [$\chi^2(1) = .053$, $p = .818$]. Twelve patients had ischaemic infarctions, 3 haemorrhagic infarctions, 1 patient had a stroke due to vasculitis. At the time of the examination patients were in a sub-acute to chronic state (1–52 months post-stroke, $M = 14.9$, $SD = 16.3$). For an overview on groups' demographics and individual clinical characteristics of the patient group see also Tables 1 and 2. Patients were recruited from three different neurorehabilitation clinics (University Hospital Bern, Kantonsspital Luzern, and Spitalzentrum Biel). All subjects had normal or corrected to normal visual acuity and an intact central visual field of 30°. All subjects gave written informed consent prior to the experiment. Ethical approval to conduct this study was provided by the Ethical Committee of the State of Bern and the State of Luzern. The present study was conducted in accordance with the principles of the latest version of the Declaration of Helsinki.

2.2. Clinical assessments

Aphasic patients were assessed on two subtests of the Aachener Aphasia Test (Huber, Poeck, & Willmes, 1984), the Token Test and the Written Language. Willmes, Poeck, Weniger, and Huber (1980) demonstrated that the discriminative validity of these subtests is as good as the discriminative validity of the whole test battery. Apraxia was examined using the imitation subscale of the standardized test of upper limb apraxia, TULIA (Vanbellingen et al., 2010). In order to exclude confounding of language comprehension and pantomime production in severely affected patients, the pantomime subscale was not applied. Handedness was measured with the Edinburgh Handedness Inventory (Oldfield, 1971).

Table 1 – Demographic and clinical characteristics.

		Aphasics $n = 16$	Controls $n = 23$
Age (in years)	Mean	52.6	50.3[a]
	Range	34–73	23–74
Gender	Male	11	15[b]
	Female	5	8
Months	Mean	14.9	
post-onset	SD	16.3	
Token Test (errors, max 50)	Mean	22.8	
	SD	14.5	
Written Language	Mean	55.7	
(correct; max 90)	SD	32.1	
TULIA (correct; max 120)	Mean	90.9	
	SD	19.7	

Note. SD = Standard Deviation; Token Test: age-corrected error scores; Written Language: raw scores; TULIA: test of upper limb apraxia, cut-off <95.

[a] $t(37) = .459$; $p = .649$.

[b] $\chi^2(1) = .053$; $p = .818$.

2.3. Lesion mapping

Lesion mapping of imaging data was conducted using MRICron (Rorden, Karnath, & Bonilha, 2007). Magnetic resonance imaging (MRI) scans were available for 11 patients and computed tomography (CT) scans were available for the remaining five patients. For the available MRI scans, the boundary of the lesions was delineated directly on the individual MRI image for every single transversal slice. Both the scan and the lesion shape were then mapped into the Talairach space using the spatial normalization algorithm provided by SPM5 (http://www.fil.ion.ucl.ac.uk/spm/). For CT scans, lesions were mapped directly on the T1-weighted single subject template implemented in MRICron (Rorden & Brett, 2000).

2.4. Stimulus material

Stimulus material consisted of one practice video and four videos for the main experiment. All videos depicted a spontaneous dialogue between a female and a male actor. The dialogues were unscripted containing spontaneous speech and co-speech gestures. The conversational topics were daily issues (favourite dish, habitation, clothing, and sports) that did not require prior knowledge. The actors were blind to the purpose of the study. Different actors played in each video; thereby every video provided a different dialogue with a

Table 2 – Individual clinical characteristics of the patient group.

Subject	Gender	Age	Months post-onset	Aetiology	Aphasia severity	Token test	Written language	Video comprehension	TULIA
1	M	60	11.2	Isch	Mild	56	n/a	11.5	98
2	M	47	36.2	Isch	Mild	58	58	11.0	93
3	M	53	6.3	Isch	Moderate	47	n/a	10.0	101
4	M	69	1.0	Isch	Mild	73	56	9.0	100
5	F	36	52.1	Hem	Moderate	45	45	11.0	97
6	M	73	1.6	Isch	Mild	55	63	8.5	69
7	M	47	15.0	Isch	Mild	54	60	10.5	107
8	M	34	11.8	Vasc	Severe	29	34	7.0	75
9	F	40	5.0	Isch	Mild	54	62	10.0	108
10	F	67	1.3	Isch	Moderate to severe	49	39	7.5	51
11	F	46	46.1	Isch	Mild to moderate	62	53	11.0	105
12	F	51	3.1	Isch	Severe	29	34	1.0	43
13	M	38	3.9	Hem	Mild to moderate	51	61	11.5	101
14	M	67	30.5	Isch	Mild to moderate	50	68	8.5	108
15	M	70	10.9	Hem	Mild	62	55	8.5	98
16	M	42	2.5	Isch	Mild	57	61	4.0	101

Note: M = male, F = female; age; in years; aetiology: isch = ischaemic infarction of medial cerebral artery, hem = hemorrhagic infarction (parenchyma bleeding), vasc = vasculitis; Aphasia Severity (Huber et al., 1984), Token Test: *T*-values; Written Language: *T*-values; TULIA: sum score imitation subscale.

different conversational topic. The videos involved two younger and two elder couples standing at a bar table (diameter 70 cm) which regulated the distance between the actors. The scene was presented from a profile view (see Fig. 1). The actors were wearing neutral dark clothes standing in front of a white background to avoid visual distraction.

2.5. *Apparatus and analysis tools*

The videos were presented on a 22″ monitor with a resolution of 1680 × 1050 pixels, 32 bit colour-depth, a refresh rate of 60 Hz, and an integrated infrared eye-tracker (RED, SensoMotoric Instruments GmbH, Teltow, Germany). The RED system is developed for contact-free measurement of eye movements with automatic head-movement compensation. A major advantage of the RED is that subjects need no fixed head rest. The eye-tracking system is characterized by a sampling rate (temporal resolution) of 250 Hz, a spatial resolution of .03° and a gaze position accuracy of .4°, mainly depending on individual calibration precision.

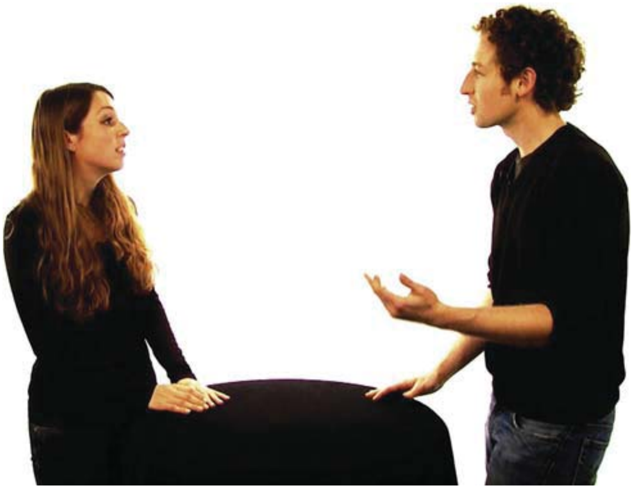

Fig. 1 – Each video depicted a different dialogue between two different actors standing at a bar table.

The eye movement recordings were pre-processed with the BeGaze™ analysis software (SensoMotoric Instruments GmbH, Teltow, Germany). Separate dynamic ROIs were defined for the hands, the face, and the body of each actor. Fixation detection threshold was set at minimal duration of 100 msec and a maximal dispersion of 100 pixels. Only fixations of the right eye were included for the analysis.

The presence of co-speech gestures was defined by the duration of their occurrence over the time course of the video using the event logging software Observer XT 10 (Noldus Information Technology bv, The Netherlands). This software allows the continuous and instantaneous sampling of behavioural video data.

The voice of the individual actors was filtered manually from the extracted sound-files of the video stimuli. Separate wav-files that now contained only the voice activity of a single actor were stored.

Furthermore, pre-processed eye movement recordings were connected with event-correlated behavioural data (co-speech gesture presence and voice activity of the actors) in Matlab 7.8.0.347 (Mathworks Inc., Natick MA). For every fixation the presence of a co-speech gesture (present or absent), the gaze direction (speaker or listener), and ROI (hands, face, or body) was defined.

2.6. *Experimental procedure*

Subjects were seated in front of the monitor, at an operating distance between 60 cm and 80 cm, their mid-sagittal plane being aligned with the middle of the screen. They were instructed to follow attentively the videos since they had to answer content-related questions after each video.

Prior to the main procedure, subjects could familiarize with the setting during a practice run which had the same structure as the following experimental trials. The main procedure consisted of four trials of video presentation that were presented in a random order. Each trial started with a 9-point calibration procedure. If gaze accuracy was sufficient (within 1° visual angle on *x*- and *y*-coordinates), the experimenter started the trial. Prior to the video stimulus, a blank screen was presented during a random interval of 1000–4000 msec followed by a fixation cross for 1000 msec. The video stimulus was presented for 2 min followed by another blank screen lasting another 2000 msec. At the end of each trial, the content-related comprehension task was performed. The experimental procedure lasted between 20 and 30 min.

2.7. Video comprehension

The aim of the video comprehension task was to verify that the subjects followed the videos attentively and to provide a general indicator of comprehension of the content. The task included 12 questions related to the content of the videos (three per video). For each video, one of the three questions was a global question about the topic, and the other two were specific questions about the contents conveyed by the two actors (one question per actor).

Each question consisted of three statements (one correct and two incorrect) that were presented individually on a pad (one statement per sheet of paper). The global question of the video comprised a correct target statement (e.g., "the woman and the man are talking about eating"), a semantically related incorrect statement (e.g., "the woman and the man are talking about drinking"), and an incorrect, unrelated statement (e.g., "the woman and the man are talking about cleaning"). The specific questions comprised one correct statement (e.g., "the man bought a new jacket") and two incorrect statements. Incorrect statements contained information that was related to the wrong actor (e.g., "the woman bought a new jacket"); semantically related but not mentioned in the video (e.g., "the man bought new shoes"); or the opposite of the video content (e.g., "the man likes to buy new clothes", in fact the man expressed his disapproval). The syntax of the statements was kept as simple as possible, with a canonical subject-verb-object (SVO) structure.

The questions were presented in a predefined order: at the beginning the global question, followed by an intermixed order of statements belonging to the specific questions about the male and the female actor.

The statements of every question were presented to the subjects in a bimodal way: in a written form (i.e., on the sheet of paper) and orally (i.e., the statements were read out by the experimenter). Subjects were instructed to judge whether each statement was correct or not, either by responding verbally or by pointing to a yes- or no-scale which was printed directly below the written form of the statements.

The score of each question was calculated on the individual responses on the corresponding statements. In order to reduce guessing probability, we adapted a scoring method known as k-prim principle (Weih et al., 2009): 3 correctly judged statements out of 3 = 1 point, 2 correctly judged statements out of 3 = .5 point, 1 or 0 correctly judged statements out of 3 = 0 point. Subjects could thus reach a maximum score of 3 points per video and 12 points throughout the whole experiment (i.e., 4 videos).

2.8. Data analysis

In a first step, pre-processed eye movement recordings were extracted from BeGaze™ analysis software for the processing with Matlab. The dataset contained now fixations on the hands, the face, or the body of the female or the male actor.

Further, the presence of co-speech gestures was rated video frame by video frame using the Observer XT 10. Co-speech gestures were rated during the stroke phase of a speech-accompanying gesture unit. The stroke phase is the main phase of a gesture unit when the movement excursion is closest to its peak (Kendon, 2004). Gesture presence was stored in a binary vector (1 = gesture, 0 = no gesture).

In addition, the speaking and the listening actor were defined over the time course of the video stimuli. According to the procedure described by Heldner and Edlund (2010), pauses (period of silence within a speaker's utterance), gaps (periods of silence between speaker changes), and overlaps (between-speaker overlaps, and within-speaker overlaps) in dyadic conversations were defined. For each video the extracted sound file was manually filtered for the voice of each actor and stored in two separate binary vectors (1 = speech, 0 = silence) that now contained only the voice activity of a single actor. The end of every gap and the start of every between-speaker overlap was defined as a point of turn during the conversation. The speaker was defined corresponding with the turn holder from turn to turn (see Fig. 2).

Finally, behavioural measures from the video stimuli about speech and co-speech gesture presence were connected with the beginning of a fixation in Matlab. For every fixation the presence of a co-speech gesture (present or absent), the gaze direction (speaker or listener), and ROI (hands, face, or body) was now defined.

Statistical analysis of behavioural and eye tracking data was conducted with IBM Statistics SPSS 21. The average duration of individual visual fixations (mean fixation duration) and summed (cumulative fixation duration) fixation duration were calculated as dependent variables. Cumulative fixation duration represents the overall time spent looking at a specific location. Statistical analysis consisted of separate mixed-design analyses of variance (ANOVA) for the dependent variables (cumulative fixation duration and mean fixation duration) and included the between-subject factor group (aphasic patients and healthy controls) and the within subject factors co-speech gesture (present and absent), gaze direction (speaker and listener), and ROI (face, hands, and hands). For the within subject factor co-speech gesture, cumulative fixation duration was weighted for the proportion of gesture presence over the video duration. For post hoc analyses, Bonferroni-corrected *t*-tests were applied. In addition, cumulative fixation duration was correlated with scores of the comprehension task, the subtests of the AAT (Token Test and

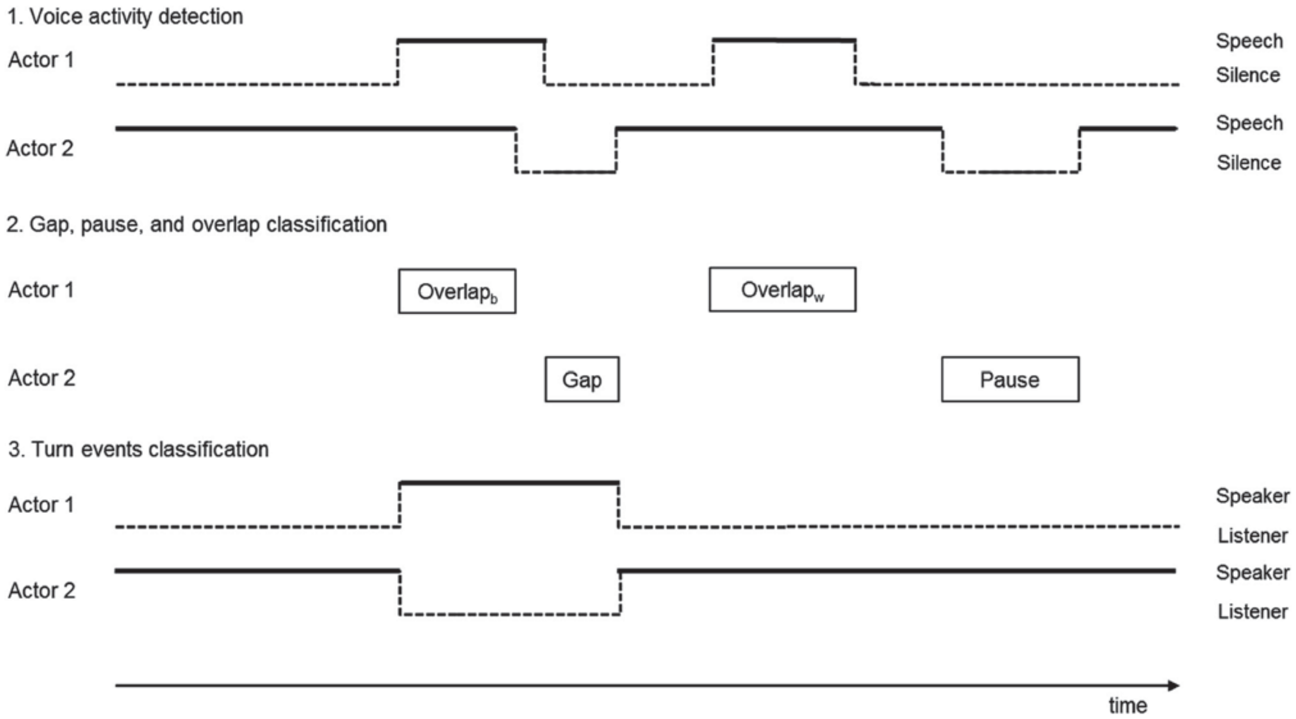

Fig. 2 – Pauses, gaps, and overlaps (Overlap$_b$ = between-speaker; Overlap$_w$ = within-speaker) are defined according to Heldner and Edlund (2010). The end of every gap and the start of every between-speaker overlap correspond to a point of turn during the conversation. The speaker and the listener are defined from turn to turn, respectively.

Written Language), and the TULIA. Significance level was set at .05 (1-tailed). Greenhouse-Geisser criterion was applied to correct for variance inhomogeneity.

3. Results

3.1. *Behavioural data*

As expected, aphasic patients ($M_{Patients} = 9.30$, $SD_{Patients} = 2.06$; $M_{Controls} = 11.37$, $SD_{Controls} = .57$) showed significantly reduced video comprehension scores [$t(36) = -4.588$, $p < .001$, 2-tailed] in comparison with healthy controls. According to the imitation subscale of the TULIA ($M = 90.9$, $SD = 19.7$), five out of 16 patients could be classified (score <95) with co-morbid apraxia.

3.2. *Analysis of fixations*

The analysis for the cumulative fixation duration revealed main effects for the factors gaze direction [$F(1,37) = 1227.623$, $p < .001$] and ROI [$F(1.07, 39.65) = 1404.228$, $p < .001$], and a gaze direction × ROI [$F(1.04, 38.48) = 846.781$, $p < .001$] interaction, indicating that subjects predominantly looked at the speaker's face.

More interestingly, we found a main effect of co-speech gesture [$F(1, 37) = 4.408$, $p = .043$], a co-speech gesture × ROI interaction [$F(1.12, 41.59) = 32.928$, $p < .001$], and a trend for a co-speech gesture × gaze direction × ROI interaction [$F(1.07, 39.42) = 3.477$, $p = .067$]. Post hoc analyses indicate that the presence of a co-speech gesture encouraged subjects to look more at the hands of the speaking actor [$t(74) = 4241.200$, $p = .010$] (Fig. 3A) and less at the listener's face [$t(74) = 14962.000$, $p < .001$] (Fig. 3D). Besides, there were a significant main effect of group [$F(1, 37) = 5.850$, $p = .021$], and a gaze direction × ROI × group [$F(2, 74) = 5.690$, $p = .005$] interaction. Aphasic patients showed reduced cumulative fixation duration on the speaker's face (Fig. 3B). Interestingly, there was no comparable between-group effect for the listener's face (Fig. 3D). Furthermore, the analysis did not reveal any significant co-speech gesture × group interaction, indicating independent effects of group and co-speech gesture.

The analysis of the mean fixation duration revealed significant main effects of co-speech gesture [$F(1, 37) = 11.082$, $p = .002$], gaze direction [$F(1, 37) = 96.661$, $p < .001$], and ROI [$F(1.36, 50.34) = 166.716$, $p < .001$]. There was a significant co-speech gesture × gaze direction interaction [$F(1, 37) = 8.813$, $p = .005$]. This means that subjects fixated longer the speaker if a co-speech gesture was present. This is supported by a significant co-speech gesture × ROI interaction [$F(1.67, 62.05) = 11.312$, $p = .001$]. Post hoc tests revealed that during the presence of a co-speech gesture subjects fixated longer the hands of the speaker [$t(74) = 300.060$, $p = .006$] (Fig. 4A).

Correlation analyses calculated for the patient group showed that visual exploration behaviour did not correlate with the score of the video comprehension task, the subtests of the AAT (Token Test and Written Language), and the imitation subscale of the TULIA. Therefore, an overall index of language impairment built from the sum of the standardized values (z-scores) of the video comprehension task and the AAT subtests was correlated with cumulative fixation duration of the speaker's face revealing a trend for a

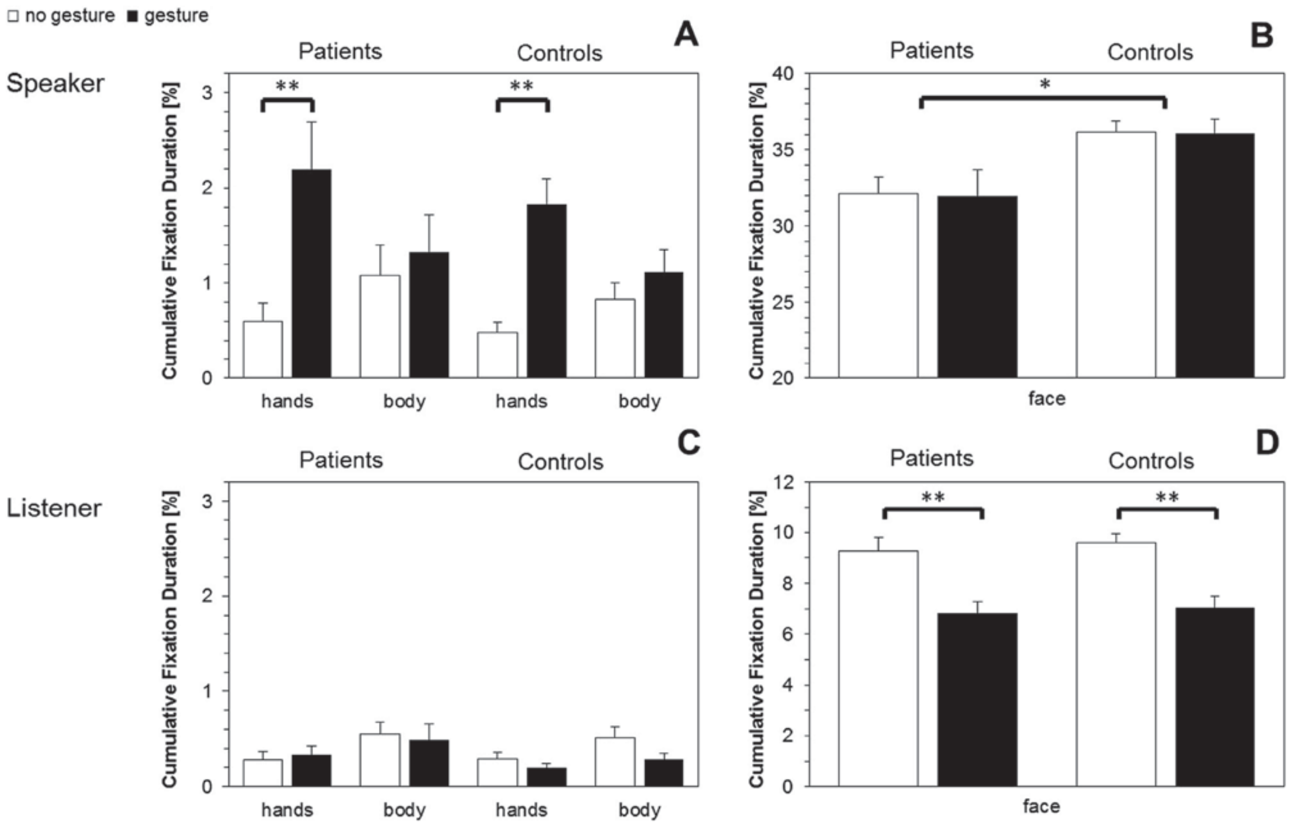

Fig. 3 — A. It demonstrates the significant increase of cumulative fixation duration on the speaker's hands if a co-speech gesture is present. B. It depicts that aphasic patients show significantly reduced cumulative fixation duration on the speaker's face. C. It illustrates no effect of co-speech gesture presence on the fixation duration of the hands and the body of the listener. D. It reveals the significant decrease of cumulative fixation duration on the listener's face if a co-speech gesture is present. Asterisks depict significant post hoc tests (*p < .05, **p < .01).

significant relation ($r_s = .374$, $p = .077$). This might imply that patients with more severe language impairments and worse video comprehension scores fixated the speaker's face less.

3.3. *Lesion analysis*

The overlay of the patients' individual cerebral lesions is shown in Fig. 5. The mean lesion volume was 96.14 cm^3 ($SD = 17.00$ cm^3). The analysis indicates a maximum overlap in the posterior superior temporal lobe (Talairach coordinates; $x = -34$, $y = -44$, $z = 10$) (see Fig. 5). However, a voxel based lesion symptom analysis including cumulative fixation duration on the speaker's face as predictor did not reach the level of significance in the *Brunner Munzel* test, probably due to the small sample size.

4. Discussion

The present study investigated the perception of video-based dyadic dialogues in mildly to severely affected aphasic patients and in healthy controls. On this account, visual exploration was measured and the influence of co-speech gestures on the fixation of dynamic ROIs (face, hand, and body of both the speaker and the listener) was analysed. It is important to consider that the dialogues in the study contained spontaneous speech and co-speech gestures, since the dialogues were unscripted and the actors were blind to the purpose of the study. The main findings are that co-speech gestures influence gaze direction and that aphasic patients fixate less the speaker's face. First, we discuss the implication that co-speech gestures guide the observer's attention throughout the dialogue. Further, we present two alternative interpretations for reduced face exploration in aphasic patients; an underlying semantic processing deficit and an audio-visual integration deficit.

4.1. *Findings in healthy control subjects*

We found that healthy controls mainly explore the speaker's face, while only a minor proportion of the cumulative fixation duration is directed towards the actors' hands. These findings are in line with previous research conducted in healthy subjects where during cartoon retelling the speaker's face is much more fixated than the gestures (Beattie et al., 2010; Gullberg & Holmqvist, 1999, 2006). People are looking at the face of their interlocutors, because eye contact plays a significant role in everyday interaction. For instance, it improves the accuracy and the efficiency in dyadic conversation during cooperative tasks (Clark & Krych, 2004).

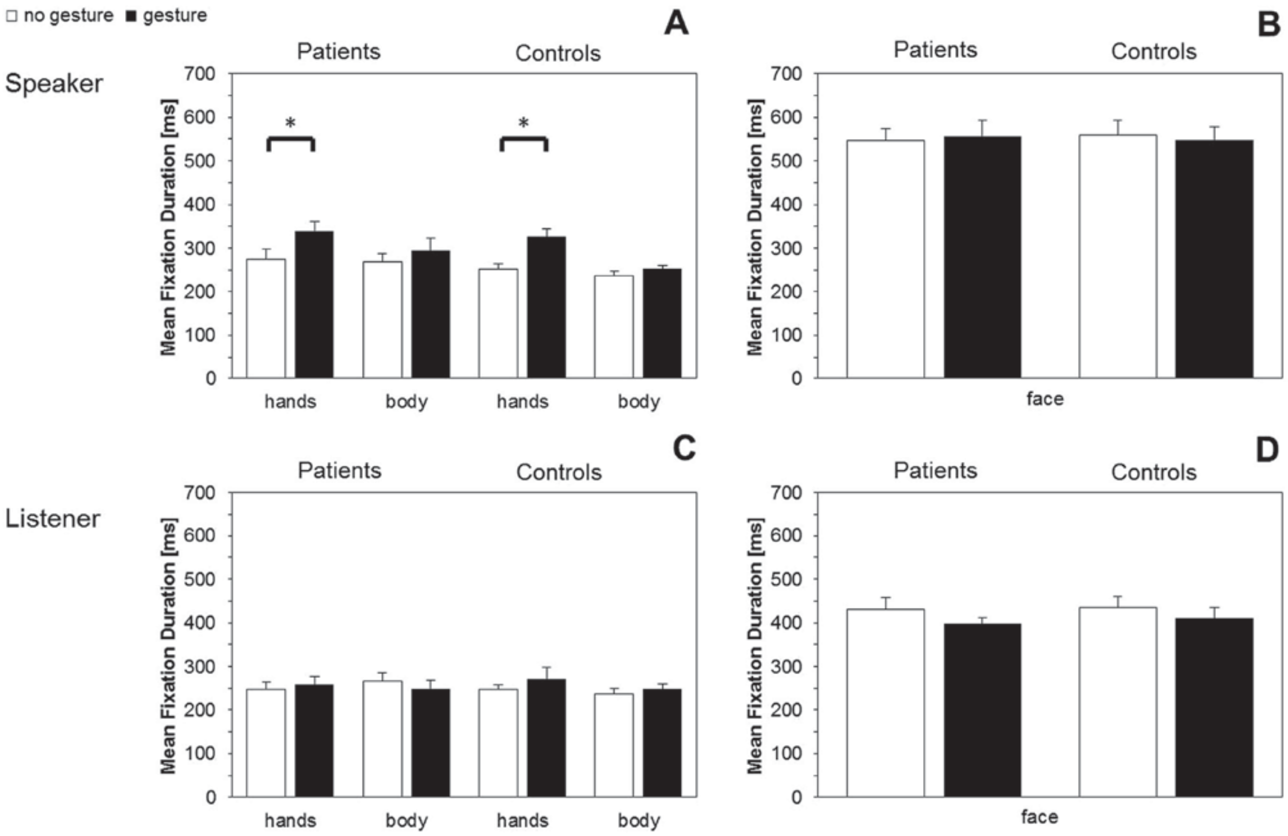

Fig. 4 – A. It demonstrates the significant increase of mean fixation duration on the speaker's hands if a co-speech gesture is present. B. It depicts equal mean fixation duration on the speaker's face in aphasic patients and healthy controls. C. It illustrates no effect of co-speech gesture presence on the hands and the body of the listener. D. It displays no effect of co-speech gesture on the listener. Asterisks depict significant post hoc tests (*$p < .05$).

More importantly, we found evidence for our first hypothesis that co-speech gestures modulate gaze direction of the observer towards the speaking actor in the video. The presence of co-speech gestures enhanced cumulative fixation duration and mean fixation duration on the speaker, and simultaneously reduced cumulative fixation duration on the listener. In particular, healthy subjects fixated longer the speaker's hands and attended less the listener's face if a co-speech gesture was present. Duncan (1972) classified gestures as one of six behavioural cues that serve as a turn-

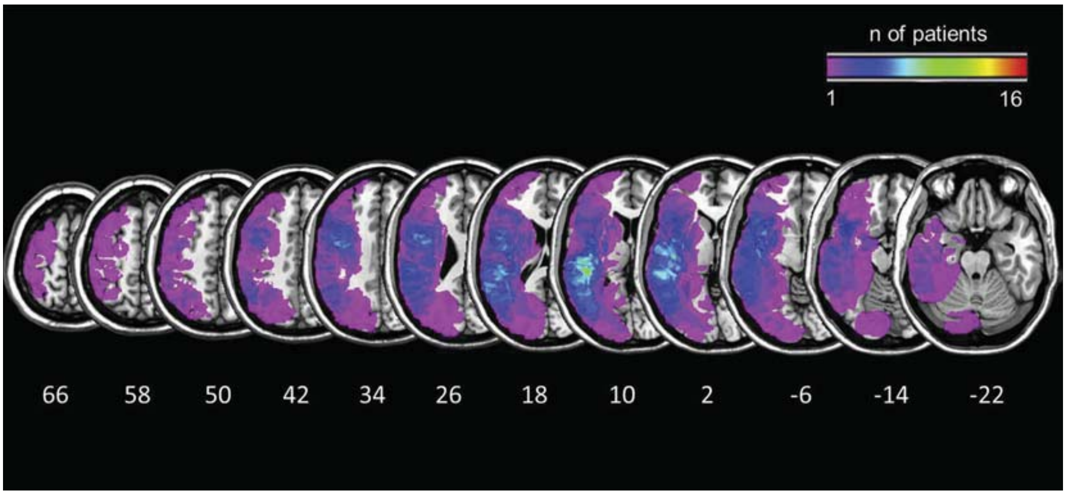

Fig. 5 – Overlap map showing the brain lesions of the 16 aphasic patients. The z-position of each axial slice in the Talairach stereotaxic space is presented at the bottom of the figure.

yielding signal. It might well be that also non-participating observers detect this signal and then direct their gaze towards the new speaker.

In addition, previous studies demonstrated that sentences presented together with a representational gesture were better recalled (Cohen & Otterbein, 1992; Feyereisen, 2006). Co-speech gestures might activate a motor image that matches an underlying representation of the word's semantics (Macedonia, Müller, & Friederici, 2011). This in turn could lead to a deeper encoding because of multi-modal representation in memory (Feyereisen, 2006). If co-speech gestures serve as a signal to identify the turn holder, they do not only lead to a deeper encoding through multi-modal representation, they also guide our attention towards the source of semantic input. Our results indicate that co-speech gestures are signalling who is holding the turn and thus indicate where the observer should look at.

4.2. *Visual exploration behaviour in aphasic patients*

Aphasic patients showed a similar processing of co-speech gestures at the perceptual level as healthy controls. The presence of co-speech gestures enhanced cumulative fixation duration and mean fixation duration on the speaker's hands and reduced cumulative fixation duration on the listener's face. This implies that the perception of co-speech gestures may also help aphasic patients to identify the turn holder and to guide their attention towards the speaker. Aphasic patients did not show increased compensatory visual exploration of the face or the hand region. Thus, we found no evidence for the first two visual exploration strategies formulated in hypothesis two: Aphasic patients neither fixate the speaker's face more in order to compensate comprehension deficits by focusing on the visual speech signal, nor did they fixate the speaker's co-speech gestures more in order to compensate verbal deficits with additional nonverbal input.

Our data show that, independent of co-speech gesture presence, aphasic patients had significantly reduced cumulative fixation duration on the speaker's face. However, both groups explored the listener's face with an equal amount of cumulative fixation duration. Since the presence of articulatory movements of the speaker's face was shown to facilitate auditory speech perception (Arnal et al., 2009; van Wassenhove et al., 2005), one could assume that the reduced visual exploration of the speaker's face diminishes the facilitating effect of audio-visual speech perception in aphasic patients. Furthermore, one might argue that reduced face exploration is due to a general impairment of visual exploration or an underlying semantic processing deficit that affects visual exploration strategies. A general impairment is unlikely since we found a specific decrease of cumulative fixation duration on the speaker's face without affecting cumulative fixation duration on the listener. Yee and Sedivy (2006), as well as Hwang, Wang, and Pomplun (2011) found that eye movements in real-world scenes are guided by semantic knowledge. It is known that the semantic knowledge may be affected in aphasic patients (Jefferies & Lambon Ralph, 2006), which may result in the observed visual exploration pattern.

An alternative explanation is offered by our third suggestion formulated in hypothesis two: Aphasic patients allocate limited attentional resources more to the acoustic speech signal and devote less attention to the visual speech signal. It was shown earlier that linguistic performance in aphasic patients degrade under conditions of higher cognitive demands such as divided attention (Erickson, Goldinger, & LaPointe, 1996; LaPointe & Erickson, 1991). The observation of dyadic dialogue does not require the constant division of attention as in dual-task paradigms. It is more complex because the cognitive demands are constantly changing throughout the dialogue. The observer has to focus on the contents provided by the interlocutors and needs to monitor constantly the collaborative processes between them in order to anticipate the next turn transition. This means that the observer has to shift his/her focus of attention permanently. Moreover, the observer encounters situations of competing speech if the interlocutors' utterances are overlapping. Kewman, Yanus, and Kirsch (1988) showed that competing speech impairs the comprehension of spoken messages in brain-damaged patients. Furthermore, there are indications that aphasic patients could have a deficit to integrate visual and auditory information. Campbell et al. (1990) suggested that the left hemisphere is important for the phonological integration of audio-visual information and the right hemisphere is important for visual aspects of speech such as face processing. It might be that preserved perceptive information of the speaker's face cannot be matched with the phonological input from the auditory speech signal. Schmid and Ziegler (2006) found in an audio-visual matching task that aphasic patients showed significantly higher error rates than healthy subjects in the cross-modal matching of visible speech movements with auditory speech sounds. According to the authors, this finding implies that aphasic patients cannot exploit as much auxiliary visual information as healthy controls. The authors concluded that the integration of visual and auditory information in their patients was impaired at the latter stage of supra-modal representations. We suggest that auditory and visual information is no longer processed congruently because integration of the auditory speech signal is impaired whereas face processing is not. The incongruence between the two signals leads to an experience of interference in aphasic patients. As a consequence, aphasic patients focus on the signal that carries more information, which is the auditory speech signal.

On the assumption that aphasic patients have a deficit to integrate audio-visual information it is interesting to consider the neural underpinnings of this deficit. There is converging evidence that the superior temporal sulcus (STS), an area that is part of the perisylvian region, and which is often affected in aphasic patients, is associated with the multisensory integration of auditory and visual information (Calvert, Campbell, & Brammer, 2000; Stevenson & James, 2009). Beauchamp, Nath, and Pasalar (2010) showed reduced cross-modal integration in healthy subjects if the STS was inhibited by transcranial magnetic stimulation. The overlay of the individual cerebral lesion maps of our patients also suggests a predominant overlap in the posterior superior temporal lobe.

Moreover, it is interesting to consider that the perisylvian region is also involved in the integration of iconic gestures and speech (Holle, Obleser, Rueschemeyer, & Gunter, 2010; Straube, Green, Bromberger, & Kircher, 2011; for reviews see also Andric & Small, 2012; Marstaller & Burianová, 2014). Furthermore, the posterior STS has been reported to be part of the action observation network (Georgescu et al., 2014) and incorporated in a neural network activated in social interaction (Leube et al., 2012). In addition to that, there is evidence from studies on patients with brain lesions that the perisylvian region is a critical site for gesture processing (Kalénine, Buxbaum, & Coslett, 2010; Nelissen et al., 2010; Saygin, Dick, Wilson, Dronkers, & Bates, 2003).

However, previous studies on gesture perception do not necessarily imply that aphasic patients with posterior temporal lesions would not be able to benefit from multi-modal presentation including speech and gesture. Findings from recent studies suggest that the use of gestures can improve naming abilities by facilitating lexical access (Göksun, Lehet, Malykhina, & Chatterjee, 2013; Marshall et al., 2012). Moreover, Records (1994) showed earlier that aphasic patients relied more on visual information provided by referential gestures if auditory information was more ambiguous. On the other hand, Cocks, Sautin, Kita, Morgan, and Zlotowitz (2009) found that the multi-modal gain of a bimodal presentation (gesture and speech) was reduced in a case of aphasia compared to a healthy control group.

4.3. *Conclusion*

In this study we investigated co-speech gesture perception in aphasic patients in a dyadic dialogue condition. We show that co-speech gestures attract only a minor portion of attention during the observation of dyadic dialogues in aphasic patients as well as in healthy controls. However, co-speech gestures seem to guide the observer's attention in both groups towards the speaker, the source of semantic input. This might indirectly facilitate deeper encoding through multi-modal representation as it was suggested in previous studies. Another finding from the present work is that aphasic patients spent less time fixating the speaker's face, probably due to an underlying sematic processing deficit or a deficit in processing audio-visual information causing aphasic patients to avoid interference between the visual and the auditory speech signal.

Acknowledgements

This study was supported by the Swiss National Science Foundation [Grant no. 320030_138532/1]. We also would like to thank Sandra Perny, Susanne Zürrer, Julia Renggli, Marianne Tschirren, Corina Wyss, Carmen Schmid, Gabriella Steiner, Monica Koenig-Bruhin Nicole Williams, Reto Hänni, Gianni Pauciello, Silvia Burren, Andreas Löffel, Michael Schraner, Nina Kohler, Anita Mani-Luginbühl, Hans Witschi, Sarah Schaefer, Martin Zürcher, and Michael Rath for their assistance.

REFERENCES

Andric, M., & Small, S. L. (2012). Gesture's neural language. *Frontiers in Psychology, 3*, 99. Retrieved from http://www.frontiersin.org.

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience, 29*(43), 13445–13453.

Beattie, G., Webster, K., & Ross, J. (2010). The fixation and processing of the iconic gestures that accompany talk. *Journal of Language and Social Psychology, 29*(2), 194–213.

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *Journal of Neuroscience, 30*(7), 2414–2417.

Behrmann, M., & Penn, C. (1984). Non-verbal communication of aphasic patients. *International Journal of Language and Communication Disorders, 19*(2), 155–168.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*(11), 649–657.

Campbell, R., Garwood, J., Franklin, S., Howard, D., Landis, T., & Regard, M. (1990). Neuropsychological studies of auditory visual fusion illusions – 4 case-studies and their implications. *Neuropsychologia, 28*(8), 787–802.

Cicone, M., Wapner, W., Foldi, N., Zurif, E., & Gardner, H. (1979). The relation between gesture and language in aphasic communication. *Brain and Language, 8*(3), 324–349.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*(1), 62–81.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1–39.

Cocks, N., Sautin, L., Kita, S., Morgan, G., & Zlotowitz, S. (2009). Gesture and speech integration: an exploratory study of a man with aphasia. *International Journal of Language & Communication Disorders, 44*(5), 795–804.

Cohen, R. L., & Otterbein, N. (1992). The mnemonic effect of speech gestures: pantomimic and non-pantomimic gestures compared. *European Journal of Cognitive Psychology, 4*(2), 113–139.

Cook, S. W., Duffy, R. G., & Fenn, K. M. (2013). Consolidation and transfer of learning after observing hand gesture. *Child Development, 84*(6), 1863–1871.

Damasio, A. R. (1992). Aphasia. *The New England Journal of Medicine, 326*(8), 531–539.

Duffy, R. J., Duffy, J. R., & Pearson, K. L. (1975). Pantomime recognition in aphasics. *Journal of Speech & Hearing Research, 18*(1), 115–132.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23*(2), 283–292.

Erickson, R. J., Goldinger, S. D., & LaPointe, L. L. (1996). Auditory vigilance in aphasic individuals: detecting nonlinguistic stimuli with full or divided attention. *Brain and Cognition, 30*(2), 244–253.

Feyereisen, P. (2006). Further investigation on the mnemonic effect of gestures: their meaning matters. *European Journal of Cognitive Psychology, 18*(2), 185–205.

Georgescu, A. L., Kuzmanovic, B., Santos, N. S., Tepest, R., Bente, G., Tittgemeyer, M., et al. (2014). Perceiving nonverbal behavior: neural correlates of processing movement fluency and contingency in dyadic interactions. *Human Brain Mapping, 35*(4), 1362–1378.

Glosser, G., Wiener, M., & Kaplan, E. (1986). Communicative gestures in aphasia. *Brain and Language, 27*(2), 345–359.

Göksun, T., Lehet, M., Malykhina, K., & Chatterjee, A. (2013). Naming and gesturing spatial relations: evidence from focal brain-injured individuals. *Neuropsychologia, 51*(8), 1518–1527.

Gullberg, M., & Holmqvist, K. (1999). Keeping an eye on gestures: visual perception of gestures in face-to-face communication. *Pragmatics & Cognition, 7*(1), 35–63.

Gullberg, M., & Holmqvist, K. (2006). What speakers do and what addressees look at: visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition, 14*(1), 53–82.

Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: eye movements and information uptake. *Journal of Nonverbal Behavior, 33*(4), 251–277.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*(4), 555–568.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Reviews of Psychology, 50*, 243–271.

Herrmann, M., Reichle, T., Lucius-Hoene, G., Wallesch, C.-W., & Johannsen-Horbach, H. (1988). Nonverbal communication as a compensative strategy for severely nonfluent aphasics? A quantitative approach. *Brain and Language, 33*(1), 41–54.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4*(4), 131–138.

Hirvenkari, L., Ruusuvuori, J., Saarinen, V. M., Kivioja, M., Perakyla, A., & Hari, R. (2013). Influence of turn-taking in a two-person conversation on the gaze of a viewer. *PLoS One, 8*(8). Retrieved form http://www.plosone.org.

Holle, H., Obleser, J., Rueschemeyer, S.-A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage, 49*(1), 875–884.

Huber, W., Poeck, K., & Willmes, K. (1984). The Aachen aphasia test. *Advances in Neurology, 42*, 291–303.

Hula, W. D., & McNeil, M. R. (2008). Models of attention and dual-task performance as explanatory constructs in aphasia. *Seminars in Speech and Language, 29*(3), 169–187.

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research, 51*(10), 1192–1205.

Jefferies, E., & Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: a case-series comparison. *Brain, 129*(8), 2132–2147.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kalénine, S., Buxbaum, L. J., & Coslett, H. B. (2010). Critical brain regions for action recognition: lesion symptom mapping in left hemisphere stroke. *Brain, 133*(11), 3269–3280.

Kendon, A. (Ed.). (2004). *Gesture: Visible action as utterance*. United Kingdom: Cambridge University Press.

Kewman, D. G., Yanus, B., & Kirsch, N. (1988). Assessment of distractibility in auditory comprehension after traumatic brain injury. *Brain Injury, 2*(2), 131–137.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57*(3), 396–414.

Krauss, R., & Hadar, U. (1999). The role of speech-related arm/hand gesture in word retrieval. In R. Campbell, & L. Messing (Eds.), *Gesture, speech and sign* (pp. 93–116). Oxford, UK: Oxford University Press.

Lanyon, L., & Rose, M. L. (2009). Do the hands have it? The facilitation effects of arm and hand gesture on word retrieval in aphasia. *Aphasiology, 23*(7–8), 809–822.

LaPointe, L. L., & Erickson, R. J. (1991). Auditory vigilance during divided task attention in aphasic individuals. *Aphasiology, 5*(6), 511–520.

Leube, D., Straube, B., Green, A., Blumel, I., Prinz, S., Schlotterbeck, P., et al. (2012). A possible brain network for representation of cooperative behavior and its implications for the psychopathology of schizophrenia. *Neuropsychobiology, 66*(1), 24–32.

Lott, P. (1999). *Gesture and aphasia*. Bern; Berlin; Bruxelles; Frankfurt am Main; New York; Wien: Lang.

Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping, 32*(6), 982–998.

Marshall, J., Best, W., Cocks, N., Cruice, M., Pring, T., Bulcock, G., et al. (2012). Gesture and naming therapy for people with severe aphasia: a group study. *Journal of Speech Language and Hearing Research, 55*(3), 726–738.

Marstaller, L., & Burianová, H. (2014). The multisensory perception of co-speech gestures – a review and meta-analysis of neuroimaging studies. *Journal of Neurolinguistics, 30*(0), 69–77.

Nelissen, N., Pazzaglia, M., Vandenbulcke, M., Sunaert, S., Fannes, K., Dupont, P., et al. (2010). Gesture discrimination in primary progressive aphasia: the intersection between gesture and language processing pathways. *Journal of Neuroscience, 30*(18), 6334–6341.

Nobe, S., Hayamizu, S., Hasegawa, O., & Takahashi, H. (2000). Hand gestures of an anthropomorphic agent: listeners' eye fixation and comprehension. *Cognitive Studies, 7*(1), 86–92.

Ochipa, C., & Gonzalez Rothi, L. J. (2000). Limb apraxia. *Seminars in Neurology, 20*(4), 471–478.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*(1), 97–113.

Records, N. L. (1994). A measure of the contribution of a gesture to the perception of speech in listeners with aphasia. *Journal of Speech and Hearing Research, 37*(5), 1086–1099.

Rimé, B., & Schiaratura, L. (1991). Gesture and speech. In R. S. Feldman, & B. Rimé (Eds.), *Fundamentals of nonverbal behavior* (pp. 239–281). Paris, France: Editions de la Maison des Sciences de l'Homme.

Rorden, C., & Brett, M. (2000). Stereotaxic display of brain lesions. *Behavioural Neurology, 12*(4), 191–200.

Rorden, C., Karnath, H. O., & Bonilha, L. (2007). Improving lesion-symptom mapping. *Journal of Cognitive Neuroscience, 19*(7), 1081–1088.

Rousseaux, M., Daveluy, W., & Kozlowski, O. (2010). Communication in conversation in stroke patients. *Journal of Neurology, 257*(7), 1099–1107.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735.

Saygin, A. P., Dick, F., Wilson, S., Dronkers, N. F., & Bates, E. (2003). Neural resources for processing language and environmental sounds: evidence from aphasia. *Brain, 126*(4), 928–945.

Schmid, G., & Ziegler, W. (2006). Audio-visual matching of speech and non-speech oral gestures in patients with aphasia and apraxia of speech. *Neuropsychologia, 44*(4), 546–555.

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition. *NeuroImage, 44*(3), 1210–1223.

Straube, B., Green, A., Bromberger, B., & Kircher, T. (2011). The differentiation of iconic and metaphoric gestures: common and unique integration processes. *Human Brain Mapping, 32*(4), 520–533.

Vanbellingen, T., Kersten, B., Van Hemelrijk, B., Van de Winckel, A., Bertschi, M., Müri, R., et al. (2010). Comprehensive assessment of gesture production: a new test of upper limb apraxia (TULIA). *European Journal of Neurology, 17*(1), 59–66.

Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2000). Why conversational agents should catch the eye. In G. Szwillus, & T. Turner (Eds.), *Conference on human factors in computing systems (CHI) 2000* (pp. 257–258). The Hague, The Netherlands: ACM Press.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181–1186.

Weih, M., Harms, D., Rauch, C., Segarra, L., Reulbach, U., Degirmenci, U., et al. (2009). Quality improvement of multiple choice examinations. In psychiatry, psychosomatic medicine, psychotherapy, and neurology. *Nervenarzt, 80*(3), 324–328.

Willmes, K., Poeck, K., Weniger, D., & Huber, W. (1980). The aachener aphasia test – differential validity. *Nervenarzt, 51*(9), 553–560.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology-Learning Memory and Cognition, 32*(1), 1–14.

Youse, K. M., Cienkowski, K. M., & Coelho, C. A. (2004). Auditory-visual speech perception in an adult with aphasia. *Brain Injury, 18*(8), 825–834.