

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° attribué par la

bibliothèque

|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|

THESE

pour obtenir le grade de

DOCTEUR DE L'INP Grenoble

Spécialité : Signal, Image, Parole, Télécoms

préparée au : Département de Parole et Cognition du laboratoire GIPSA-Lab

dans le cadre de **l'Ecole Doctorale** : Electronique, Electrotechnique, Automatique, Traitement du Signal

présentée et soutenue publiquement

par

Stephan Raidt

le 2 Avril 2008

TITRE

Gaze and face-to-face communication between a human speaker and an animated conversational agent – Mutual attention and multimodal deixis

Regard et communication face-à-face entre un locuteur humain et un agent conversationnel animé. Attention mutuelle et monstration multimodale

DIRECTEUR DE THESE

Gérard Bailly

CO-DIRECTEUR DE THESE

Laurent Bonnaud

JURY

M.	James Crowley	, Président
Mme.	Catherine Pelachaud	, Rapporteur
M.	Dirk Heylen	, Rapporteur
M.	Gérard Bailly	, Directeur de thèse
M.	Laurent Bonnaud	, Co-encadrant

Mme. Noëlle Carbonell, Examinatrice

ABSTRACT

In the context of synthetic generation and decoding of linguistic information, not only the audible component but also the visual component of speech conveys valuable information. We address gaze as an important modality to enhance speech and to convey additional information. Gaze is an important deictic gesture as well as it plays various roles in the organization of dialogue and social interaction.

In a first experiment, we investigated how the gaze of a talking head can be used as a deictic gesture in an on-screen search and retrieval task. We found that such gestures are appropriate to reduce processing time as well as cognitive load. Multimodal gestures incorporating speech in a coherent way showed to be more efficient than only visual gestures.

In a second experiment, we investigated the relations between the gaze of a target subject and different elements of conversational interaction. We defined different stages in the dialogic exchange of information and found that these are related to the variations in the measured gaze behavior. Based on the observed characteristics we propose a model to control the gaze of an embodied conversational agent in close dyadic interaction.

Dans le contexte de la génération synthétique et le décodage d'informations linguistiques, non seulement le composant auditif mais aussi le composant visuel de la parole transmettent de l'information précieuse. Nous étudions le regard en tant qu'élément crucial pour enrichir la parole et fournir des informations supplémentaires. Le regard est un geste déictique très important, ainsi il influence de manières variées l'organisation du dialogue et de l'interaction sociale.

Dans une première expérience nous étudions comment le regard d'une tête parlante peut être employé comme geste déictique dans un jeu de recherche et sélection sur un écran d'ordinateur. Nous avons trouvé que ces gestes sont capables de réduire le temps de réaction ainsi que la charge cognitive. Cet effet est accentué quand le geste est de caractère multimodal, utilisant la parole de manière adaptée.

Dans une deuxième expérience nous avons étudié le rapport entre le regard d'un sujet cible et les différents éléments d'une interaction dialogique. Nous avons défini différents segments dans l'échange d'information dialogique et ont trouvé qu'ils sont liés aux variations du comportement du regard mesuré. Basé sur ces résultats nous proposons un modèle pour le contrôle du regard d'un agent conversationnel animé dans l'interaction face-à-face.

CONTENTS

ABSTRACT.....	3
CONTENTS.....	5
LIST OF FIGURES.....	9
GLOSSARY.....	15
INTRODUCTION.....	17
1 STATE OF THE ART.....	21
1.1 THE PHYSIOLOGY OF THE EYE.....	21
1.1.1 <i>The anatomy of the eye.....</i>	21
1.1.2 <i>Eye movements and blinks.....</i>	23
1.1.2.1 Saccades.....	23
1.1.2.2 Gaze shifts as a combined movement of eyes and head.....	24
1.1.2.3 Smooth pursuit.....	25
1.1.2.4 Fixations.....	25
1.1.2.5 Blinks.....	25
1.2 MEASUREMENT OF GAZE DIRECTION.....	26
1.2.1 <i>Human observers.....</i>	26
1.2.2 <i>Mechanical techniques.....</i>	27
1.2.3 <i>EMG techniques.....</i>	28
1.2.4 <i>Video techniques.....</i>	28
1.3 VISUAL PERCEPTION OF EYES AND GAZE AS SPECIAL STIMULI.....	28
1.3.1 <i>Visual perception of eyes and gaze by children.....</i>	28
1.3.1.1 Gaze, speech and pointing.....	30
1.3.1.2 Gaze and joint visual attention.....	31
1.3.2 <i>Perception of gaze direction.....</i>	31
1.3.2.1 Precision of estimation of gaze direction.....	32
1.3.2.2 Important features for the estimation of gaze direction.....	38
1.3.2.3 Discussion.....	39
1.3.3 <i>Gaze direction, deixis and visual attention.....</i>	40
1.3.3.1 Deictic capacity of eye gaze.....	40
1.3.3.2 Gaze direction and visual attention.....	41
1.3.4 <i>Gaze patterns during examination of human faces.....</i>	42
1.3.5 <i>Gaze direction in text reading.....</i>	43
1.4 EYE GAZE AND SOCIAL INTERACTION.....	44
1.4.1 <i>Definition of terms concerning gaze.....</i>	45
1.4.2 <i>Eye gaze and theory of mind.....</i>	45
1.4.3 <i>Functions of gaze in conversation and dialogue.....</i>	47
1.4.3.1 Adam Kendon's pioneer work.....	47
1.4.3.2 Turn taking.....	49
1.4.3.3 Gaze and interpersonal distance.....	50
1.4.3.4 Social norms and gaze behavior.....	52
1.4.3.5 Gaze in multi-party interaction.....	53
1.4.3.6 Summarizing the function of gaze in conversation and dialogue.....	53
1.4.4 <i>Mediated interaction between humans.....</i>	53
1.4.5 <i>Interaction between humans and animated agents.....</i>	54
1.5 GAZE AS INPUT MODALITY IN HUMAN-COMPUTER INTERACTION.....	56
1.6 EMBODIED CONVERSATIONAL AGENTS.....	56
1.6.1 <i>Talking heads.....</i>	58
1.6.2 <i>Animation of Embodied Conversational Agents and Modeling of interaction.....</i>	59
1.6.3 <i>Gaze Modeling for Conversational Agents.....</i>	61
1.6.3.1 Modeling of gaze and scene perception.....	62
1.6.3.2 Modeling of gaze and social interaction.....	64
1.6.4 <i>Conversational Agents versus Robots.....</i>	67
1.6.5 <i>Evaluation and Perception of Conversational Agents.....</i>	70
1.6.6 <i>Conclusions.....</i>	70
2 GAZE DIRECTION AND ATTENTION.....	73

2.1	THE EXPERIMENTAL SETUP AND SCENARIO OF INTERACTION	73
2.1.1	<i>Setup</i>	73
2.1.2	<i>Scenario</i>	74
2.1.3	<i>Data acquisition</i>	76
2.1.4	<i>Data Processing</i>	76
2.1.5	<i>Variables of the analysis</i>	77
2.2	EXPERIMENT I: IMPACT OF FACIAL CUES	77
2.2.1	<i>Experimental conditions</i>	77
2.2.2	<i>Results</i>	77
2.3	EXPERIMENT II: IMPACT OF MULTIMODAL DEICTIC GESTURES.....	81
2.3.1	<i>Experimental conditions</i>	81
2.3.2	<i>Results</i>	82
2.4	DISCUSSION AND PERSPECTIVES	84
3	MEDIATED FACE-TO-FACE INTERACTION.....	87
3.1	SCENARIO AND SETUP	87
3.1.1	<i>Scenario</i>	87
3.1.1.1	Definition of conversational role and cognitive state	88
3.1.1.2	Distinction of cognitive states	89
3.1.1.3	Faked interaction using a pre-recorded stimulus	89
3.1.1.4	Control of interaction and dialogue to produce recurrent events	89
	Sentence building game	90
	Sentence repeating game.....	90
	Linguistic content.....	90
3.1.2	<i>Setup for mediated face-to-face interaction</i>	90
3.1.2.1	Choice of eye tracking system.....	92
3.1.2.2	Choice of technique for video linkage.....	92
3.1.2.3	Positioning of camera, angle of view and distance	93
3.1.2.4	Recording device (VHS, dvd, hard disc).....	94
	Audio-Video recording with ClearView	94
	Audio-Video hard disc recording and recording software.....	95
3.1.2.5	Synchronization of data.....	95
3.1.2.6	Consistency between displayed and recorded video.....	96
3.1.2.7	Consistency of video image and eye tracking.....	97
3.2	DATA PROCESSING.....	98
3.2.1	<i>Extraction of parameters from video signal</i>	98
3.2.2	<i>Extraction of parameters from audio signal</i>	98
3.2.3	<i>Preprocessing of gaze data to generate data with constant frequency</i>	98
3.2.4	<i>Fixation detection</i>	99
3.2.4.1	Fixation detection and visual processing.....	99
3.2.4.2	Algorithms for fixation detection	100
3.2.4.3	Choice of algorithm for fixation detection	100
3.2.4.4	Specification of parameters	100
3.2.5	<i>Assignment of regions of interest</i>	101
3.2.5.1	Distinction of regions of interest	101
3.2.5.2	Relating fixation to regions of interest	101
3.2.5.3	Tracking regions of interest.....	102
3.2.6	<i>Detection of blinks</i>	103
3.2.7	<i>Segmentation of interaction</i>	103
3.2.7.1	Automatic detection of cognitive states.....	103
3.2.7.2	Manual verification of automatic segmentation	104
3.3	STATISTICAL ANALYSIS	105
3.3.1	<i>Endogenous gaze control</i>	105
3.3.1.1	Fixation time	106
	Descriptive statistics.....	106
	Analytic statistics	108
	Discussion.....	110
3.3.1.2	Probability of fixations over ROI	111
3.3.1.3	Duration of fixation	113
3.3.1.4	Blinks	114
	Random occurrence of blinks.....	115
	Constant frequency of blinks.....	118
	Discussion.....	122
3.3.2	<i>Exogenous control</i>	123

3.3.2.1	Response to eye-directed gaze.....	123
3.3.2.2	Mutual Gaze.....	125
3.3.3	<i>Comparison of gaze behavior in live and faked interaction</i>	129
3.4	MODELING.....	130
3.4.1	<i>The HMM technique</i>	130
3.4.2	<i>Generating fixations</i>	131
3.4.3	<i>Generating blinks</i>	132
3.4.4	<i>Evaluation</i>	135
4	CONCLUSIONS	137
4.1	GAZE DIRECTION AND ATTENTION.....	137
4.2	MEDIATED FACE-TO-FACE INTERACTION.....	138
5	APPENDIX	140
5.1	QUESTIONNAIRE PRESENTED ON SCREEN AFTER THE CARD GAME.....	140
5.2	SETUP CHECKLIST: MEDIATE FACE-TO-FACE INTERACTION.....	141
5.3	FIXATION TIME AND FIXATION PROBABILITIES OF INTERLOCUTORS OF OUR TARGET SUBJECT.....	144
5.4	COMPARISON OF FIXATION TIME AND FIXATION PROBABILITY BETWEEN LIVE AND FAKED INTERACTION.....	150
6	REFERENCES	155
7	RESUME EN FRANÇAIS DE LA THESE	162
7.1	INTRODUCTION.....	162
7.2	ÉTAT DE L'ART – REGARD ET PERCEPTION VISUELLE.....	165
7.2.1	<i>La physiologie de l'œil</i>	165
7.2.2	<i>Mesure de la direction du regard</i>	165
7.2.3	<i>Perception visuelle des yeux et du regard</i>	165
7.2.4	<i>Le rôle du regard dans l'interaction sociale</i>	165
7.2.5	<i>Le regard comme modalité de commande dans l'interaction homme-machine</i>	166
7.2.6	<i>Agents conversationnels animés</i>	166
7.3	ATTENTION ET DIRECTION DU REGARD.....	166
7.3.1	<i>Montage expérimental et scénario d'interaction</i>	166
7.3.2	<i>Expérience I: Impact des gestes faciaux</i>	167
7.3.3	<i>Expérience II: Impact des gestes déictiques multimodaux</i>	168
7.3.4	<i>Discussion et perspectives</i>	168
7.4	INTERACTION FACE-A-FACE MEDIATISÉE.....	169
7.4.1	<i>Scénario et installation d'expérimentation</i>	169
7.4.2	<i>Acquisition et traitement de données</i>	170
7.4.3	<i>Traitement de données</i>	170
7.4.4	<i>Analyse statistique</i>	170
7.4.5	<i>Modélisation</i>	171
7.5	CONCLUSIONS.....	171
7.5.1	<i>Direction du regard et attention</i>	172
7.5.2	<i>Interaction face-a-Face mediatise</i>	173
7.6	REFERENCES DU RESUME.....	175

LIST OF FIGURES

<i>Figure 1: Anatomy of the eye</i>	21
<i>Figure 2: Diagram of the concentration of rods and cones on the retina around the fovea</i>	22
<i>Figure 3: Top and front views of the right eye and the extra-ocular muscles: medial rectus (mr), lateral rectus (lr), superior rectus (sr), inferior rectus (sr), superior oblique (so), inferior oblique (io); Sparks (2002)</i>	23
<i>Figure 4: Velocity curves observed during horizontal gaze shifts of different amplitudes, in experiments on rhesus monkeys with heads restrained (Freedman & Sparks (1997))</i>	24
<i>Figure 5: Left: Peak velocities as observed for horizontal eye movements of different amplitudes. Right: Schematic course of head and eye position of a combined movement of gaze shift (Freedman & Sparks (1997))</i>	24
<i>Figure 6: Angular distribution of microsaccades: thin line – binocular; bold line – monocular (left or right eye) (from Engbert & Kliegl (2003))</i>	25
<i>Figure 7: Left: Schematic drawing of an experimental arrangement for the gaze monitoring of two subjects (A and B) by two observers (O₁ and O₂) behind a one-way screen. Right: Pen writer protocol of the interaction of two subjects listing intervals of gaze and speech (Argyle & Cook (1976))</i>	27
<i>Figure 8: Perception of faces and face-like stimuli by children. Left: Overview over the type of stimulus needed to produce increased smiling of infants and its development over the first 9 month of life, as observed by Ahrens (1954). Right: Experimental stimuli used by Farroni, Csibra et al. (2002), along with information about viewing angles of special regions, when fixated or when in the periphery (in brackets) . (a) Direct or averted gaze (b) Images with a resolution according to the average visual acuity of newborns (c) Images with resolution according to average visual acuity of 4-month-old infants</i>	30
<i>Figure 9: Left: Schematic view from above on the experimental setup used in Gibson & Pick (1963). Right: Results in Gibson & Pick (1963), measured as frequency of positive answers to the question whether the gaze of a looker was directed towards the subject or not, when the looker adopted different head orientations relative to the observer (30° left, 0°, 30° right)</i>	32
<i>Figure 10: Left: Schematic representation of the experimental setup as used in Cline (1967). Right: Front view of target board containing the fixation targets for the looker (left) and response board on which subjects were asked to mark the estimated fixation target (right)</i>	34
<i>Figure 11: Experimental setups as used in Anstis, Mayhew et al. (1969). Left: Schematic view from above on setup with a looker fixating targets and a subject marking the estimated fixation target. Right: Similar setup but with a camera to acquire the stimulus that is presented on a TV screen turned at different angles relative to the subject</i>	34
<i>Figure 12: Charts putting the gaze direction as perceived by the subjects in relation to the actual gaze direction of the looker. The bold line represents straight head orientation of the looker, the thin line clockwise turn of the lookers head and the dashed line counter-clockwise turn of the lookers head</i>	35
<i>Figure 13: MPEG-4 animated head used in Svanfeldt, Wik et al. (2005) and three examples of eye region for accommodation to different distances</i>	36
<i>Figure 14: Left representation of angles of gaze direction of the looker relative to straight ahead. The lines delineate ranges of angles that produced the perception of eye-directed gaze at a certain percentage (measured by Chen (2002)). Right: Relation of perceived and actual angle of vertical gaze direction as measured by Anstis, Mayhew et al. (1969)</i>	37
<i>Figure 15: Examples of stimuli, with face as positive contrast representation (left) and negative contrast representation (right), both combined with positive and negative contrast for the representation of the eyes. In all four images, the head and gaze are both directed 30° to the observer's right (Ricciardelli, Baylis et al. (2000))</i>	39
<i>Figure 16: Further examples of stimuli as different combinations of head and eye orientation, as cutout of eye region only. From left to right: eyes-left with the head facing right; eyes-straight with the head facing right; straight eyes in a straight head; eyes-left in a straight face. Top row: positive contrast. Bottom row: negative contrast (Ricciardelli, Baylis et al. (2000))</i>	39
<i>Figure 17: Schematic representation of the sequence of displays used by Langton, Watt & Bruce (2000) according to the Posner paradigm. Subjects are asked to fixate the cross in the middle of the display. An image of a face is used as cue and a circle as target. The time span between the onsets of these two stimuli is the stimulus onset asynchrony (SOA)</i>	40
<i>Figure 18: Schematic diagram of the relationship between the four components proposed by Baron-Cohen's for the modeling of the human mind-reading system</i>	47

Figure 19: Example of a transcription protocol from an interaction analyzed in Kendon (1967). The visual appearance of the eyes, brows and mouth, the head position, gaze direction and the speech are annotated in intervals of 500ms. The meaning of the used symbols is given below the table.	49
Figure 20: Total amount of eye contact in ms as measured during interactions of 3 minutes at distances of 2, 6 and 10 feet (60cm, 1.8m and 3m), and different combinations of sex of subjects. The ‘confederate’ was instructed to continuously look at the subject’s eyes in order to produce mutual gaze whenever the subject looks up (Argyle & Dean (1965)).	51
Figure 21: Schematic representation of conditions as used in the experiments in Bailenson, Blascovich et al. (2001) to verify the equilibrium theory in immersive virtual environments. The middle column informs about the characteristics of the presented stimulus and its animation. The column to the right provides the corresponding visual realizations.	52
Figure 22: View of two person communicating via the ‘video tunnel’ (coupled teleprompters) as used by Garau, Slater et al. (2001).	54
Figure 23: Left: Images of the robot used by Minato, Shimada et al. (2005) in experiments on human-android interaction. Right: Repartition of regions of the face for the distinction of gaze directions.	56
Figure 24: Left: Schematic representation of the model for turn-taking proposed in Thórisson (2002), separated in three layers of different processing time, that exchange data, analyse input from scene analysis and generate output for the generation of behavior. Right: Description of the three layers mentioned in the diagram to the left, detailing the level of priority and the type of in- and output data treated in dependence of the role taken in conversation.	61
Figure 25: Left: Schematic description of the generation of gaze from saliency maps enhancing different features (motion, color, intensity, flicker, orientation, etc) of the video input used in the gaze model by Itti, Dhavale et al. (2003). Right: Chart of the model proposed in Picot, Bailly, Elisei & Raidt (2007) based on the computation of a similar saliency map but using an attention stack for switching between targets.	63
Figure 26: Comparing gaze direction as predicted by the gaze model proposed in Picot, Bailly et al. (2007) (continuous black line), with gaze as recorded from subjects viewing the same stimulus (bold gray line). Top: right – left coordinates. Bottom: up – down coordinates.	63
Figure 27: Examples of the visual appearance of the talking head used in Lee, Badler et al. (2002). Left; straight ahead gaze. Right: averted gaze.	64
Figure 28: Visual appearance of male and female avatars with straight and averted gaze, as used in Garau, Slater et al. (2001).	65
Figure 29: Avatars as used in Vinayagamoorthy, Garau et al. (2004) and Garau, Slater et al. (2003). Left: cartoon-like avatar. Middle: realistic avatars. Right: view of avatar in virtual environment.	65
Figure 30: Left: Function chart describing the behavior model of the WE-4RII robot from the Takanishi Laboratory at WASEDA University. Right: Photo of the WE-4RII robot.	68
Figure 31: Photo of ROBITA, developed at the Perceptual Computing Group at the Humanoid Robotics Institute of WASEDA University, in interaction with two persons (Matsusaka, Fujie et al. (2001)).	68
Figure 32: Left: photo of Kismet. Right: Schematic description of the active vision system implemented in Kismet – analysis of video separated in different channels; task-driven weighting of contribution from different channels for the determination of the target of attention; activation of motor system for gaze shifts towards targets of attention.	69
Figure 33: Schematic description of the finite state machine controlling the interaction of the experimental setup used for the card game experiments.	74
Figure 34: Experimental conditions: The experiment is divided into four conditions with different levels of help and guidance by the talking head. When the talking head is present it can give helpful or misleading facial cues. When the numbers on the cards are not shown (right; no digits displayed) these cues are the only possibility to locate the correct target card.	75
Figure 35: Comparing reaction time means for four pairs of conditions. From left to right: deceptive indications vs. correct indications; no assistance vs. correct indications; no assistance vs. deceptive indications; no digits displayed vs. correct indications. The x-coordinate lists the subjects whereas the digit represents the order of participation. Stars above these digits indicate statistical significance of difference between the underlying distributions at $\alpha = 0.05$. The order of subjects is sorted for increasing difference of reaction time means between the compared conditions from left to right.	78
Figure 36: Comparing the number of cards inspected during the search for the correct target card. Conditions compared and order of subjects are the same as in Figure 35.	80
Figure 37: Comparing reaction times for four pairs of conditions. From left to right: condition 2 vs. condition 3; condition 1 vs. condition 3; condition 1 vs. condition 2; condition 4 vs. condition 3. The x-coordinate lists the subjects whereas the digit represents the order of participation. Mean reaction times for each user and for each session are displayed together with the statistical significance of the underlying distributions (stars displayed at the bottom when $p < 0.05$).	83

<i>Figure 38: Comparing the number of cards inspected during the search for the correct target card. Conditions compared and order of subjects are the same as in Figure 37</i>	<i>83</i>
<i>Figure 39: Evolution of eye model: Left: modeling of eyelids as two triangles described by four vertices to realize closing of eye and blinks. Middle: measurement of shape of eyelids in relation to the position of the pupil for the development of an advanced eyelid model. Right: implementation of eyelid model in the female talking head.....</i>	<i>86</i>
<i>Figure 40: Setup for mediated face-to-face interaction of two subjects. The audio-visual appearance of a subject is captured with a microphone and a camera and presented to the other subject with loudspeakers and a computer screen. The signals are recorded along with the gaze data acquired with the eye trackers that are integrated into the screens.</i>	<i>91</i>
<i>Figure 41: Setup for mediated interaction of a subject with a prerecorded audiovisual stimulus. The stimulus is played back from a VHS audio and video recording. The played back audio signal is recorded again on a separate channel along with the audio signal of the present subject, which enables temporal alignment of the data after the experiment.</i>	<i>91</i>
<i>Figure 42: Chart of different procedures performed in connection with an experimental recording. The check using the LASER-pointer is not performed when using a pre-recorded stimulus during the faked interaction. ...</i>	<i>92</i>
<i>Figure 43: Comparison of perceived and actual target of fixation. Left: Schematic drawing of the scene for the marking of actual or estimated gaze target. Middle and right: The stars mark the fixated targets of the observer. Arrows indicate the location that the observer (person shown in picture) estimated to be the fixation target, if there is a discrepancy between his estimation and the actually fixated target.....</i>	<i>94</i>
<i>Figure 44: Stimulus presented on the screen during the synchronization procedure: left: initial image; right: indicated circular movement of gaze target to generate smooth a sequence of valid gaze data.....</i>	<i>96</i>
<i>Figure 45: Example for the matching of gaze data. The X and Y coordinates for the left (L) and right eye (R) are displayed for the data streams recorded with the two different software tools. As expected, the respective curves can be matched and superposed. 'CV' stands for the data acquired with Clearview that is represented by continuous lines. 'TET' stands for the data acquired with TETserver represented by 'X'. The synchronization event is marked with a dashed vertical line.</i>	<i>96</i>
<i>Figure 46: Extract from the graphical representation of the gaze direction of our target subject (as respondent in interaction 8). The x-coordinates of gaze direction are represented in black, y-coordinates in gray. Crosses mark gaze points in 20 ms steps on the time axes (abscissa) and in pixels on the ordinate whereas the gaze targets on the screen are represented at a resolution of 320x240 pixels. The horizontal bold lines represent detected fixations with the means of the x- and y-coordinates of the gaze points that belong to respective same fixation.</i>	<i>101</i>
<i>Figure 47: Assignment of fixations to regions of interest. The ellipses are determined by hand to delineate the fixations assigned to the ROI mouth, right eye, left eye and face. The size of the asterisks is proportional to the duration of fixations. Left: fixations of our target speaker on face of subject, Right: the interlocutor's data. ...</i>	<i>102</i>
<i>Figure 48: Multimedia annotation software ELAN® (Hellwig & Uytvanck (2004)) as used for manual verification and correction of automatic segmentation. Top: video images with superimposed fixation targets. Middle: audio signal. Bottom: annotation tiers for cognitive state, fixation and blink.</i>	<i>105</i>
<i>Figure 49: Box-plots of repartition of fixation time during an instance of a cognitive state during a given role. The box-plots are calculated from the proportion of time that was dedicated to a ROI during the instances of a cognitive state. Here the ROI face, right eye, left eye and mouth that were taken into account for the statistical test are listed.</i>	<i>107</i>
<i>Figure 50: The diagrams show a 3D representation of the repartition of mean fixation time dedicated to the different ROI in %/100. The bars represent mean values of the proportion of time that was dedicated to a ROI during the instances of a cognitive state. A detailed representation of the repartition of data is given in Figure 49. Abscise: speaking, listening, waiting, reading, pre-phonation, thinking else; ordinate: face, right eye, left eye, mouth, else (represented by initial letter).</i>	<i>108</i>
<i>Figure 51: Projection on the first discrimination planes of the groups formed by the combinations of CS and role as indicated in the diagrams with the respective abbreviations. The groups are represented as a projection on the first two main axes in the diagram to the left, and on the first and third axis in the diagram to the right. ...</i>	<i>109</i>
<i>Figure 52: The diagrams show a 3D representation of the probabilities that a ROI is fixated at least once during a CS in %/100. Abscise: speaking, listening, waiting, reading, pre-phonation, thinking else; ordinate: face, right eye, left eye, mouth, else (represented by initial letter).....</i>	<i>112</i>
<i>Figure 53: Histogram of the duration of all fixations of the reference subject of all interactions. To obtain a symmetrical distribution the natural logarithm of the values is taken. The dashed line indicates the mean, which is also given as value, along with the standard deviation and the duration in ms corresponding to the mean. .</i>	<i>113</i>
<i>Figure 54: Same as Figure 53 but grouped for the CS during which the fixations were observed.....</i>	<i>113</i>

<i>Figure 55: Same as Figure 54 but grouped for the CS and the role during which the fixations are observed. The left diagram shows the histogram for the role of initiator, the right diagram shows the histogram for the role of respondent.....</i>	<i>114</i>
<i>Figure 56 : Blink rate, calculated as total number of blinks per sum of duration of CS for a given interaction (1 ... 9) and role (left: initiator, right: respondent). Abscise: listening, waiting, speaking, pre-phonation (represented by initial letter).....</i>	<i>117</i>
<i>Figure 57: Histogram of all inter-blink intervals measured during the interactions of our target subject, represented as measured in milliseconds (left) and after transformation to the natural logarithm (right). The dashed line indicates the mean, given also as a figure in the diagram along with the standard deviation and the corresponding mean in milliseconds in the case of the logarithmic representation.</i>	<i>120</i>
<i>Figure 58: Deviation of inter-blink duration from mean, measured during the interactions of our target subject, separated for CS and role (initiator left; respondent right). Only the CS speaking, listening, waiting and pre-phonation that are common to both roles and appear with a minimum frequency are considered.</i>	<i>120</i>
<i>Figure 59: Screen capture from the graphical interface of the annotation tool ELAN® showing an example of an inter-blink interval strongly delayed relative to the mean. The concerned CS listening is hence classified as inhibiting blink.</i>	<i>121</i>
<i>Figure 60: Histogram of all inter-blink intervals measured during the interactions of our target subject that can not be associated with the start or end of a CS. They are represented as measured in milliseconds (left) and after transformation to the natural logarithm (right). The dashed line indicates the mean, given also as a figure in the diagram along with the standard deviation and the corresponding mean in milliseconds in the case of the logarithmic representation.....</i>	<i>122</i>
<i>Figure 61: Deviation of inter-blink duration from mean, measured during the interactions of our target subject, separated for CS and role. Only the CS speaking, listening, waiting and pre-phonation that are common to both roles and appear with a minimum frequency are considered. The left diagram displays all measured intervals. In the right diagram, intervals that can be associated with the start or end of a CS are excluded. Data show that listening tends to slow down blink rate whereas speaking speeds up the blink rate.</i>	<i>122</i>
<i>Figure 62 : Mean response of target subject to eye-directed gaze of interacting subjects as signal (continuous line). For every instance that a subject looked at the right or left eye of the target subject, the gaze direction of the latter was noted in 20ms steps as either directed (1) or not (0) towards a ROI (- mouth; -- right eye; - left eye). From the observations of all intervals taken into account the mean response was calculated, as the sum divided by the number of instances. The peak represents the mean signal of eye-directed gaze of the subjects.</i>	<i>124</i>
<i>Figure 63: Mean response of interacting subjects to eye-directed gaze of target subject as signal (continuous line). For every instance that the target subject looked at the right or left eye of the interlocutor, the gaze direction of the latter was noted in 20ms steps as either directed (1) or not (0) towards a ROI (- mouth; -- right eye; - left eye). From the observations of all intervals taken into account the mean response was calculated, as the sum divided by the number instances. The peak represents the mean signal of eye-directed gaze of the target subject.</i>	<i>125</i>
<i>Figure 64: Percentage of mutual gaze. The durations of intervals during which both subjects direct their gaze to the others eyes are added up separated for CS and role and divided by the total duration of the respective CS. The diagram on the left shows the results for the CS speaking, listening, waiting and pre-phonation with the target subject as initiator. The diagram to the right shows the corresponding CS with the target subject as respondent.....</i>	<i>127</i>
<i>Figure 65: Percentage of the target subject's gaze directed towards the eyes of the interlocutor separated for CS and role.....</i>	<i>127</i>
<i>Figure 66: Percentage of the interlocutors' gaze directed towards the eyes of the target subject. The considered intervals and their separation for CS and role corresponds to the segmentation and labeling of the target subject's data.</i>	<i>128</i>
<i>Figure 67: Percentage of mutual gaze relative to amount of eye-directed gaze of the target subject, separated for role. The amount of mutual gaze observed during the CS speaking, listening, waiting and pre-phonation of the target subject is divided by the amount of eye-directed gaze of the target subject observed during these intervals.</i>	<i>128</i>
<i>Figure 68: Percentage of mutual gaze relative to amount of eye-directed gaze of the interlocutors, separated for role. The amount of mutual gaze observed during the CS speaking, listening, waiting and pre-phonation of the target subject is divided by the amount of eye-directed gaze of the interlocutors observed during the respectively inspected intervals.....</i>	<i>128</i>
<i>Figure 69: Mean distributions of fixation time in %/100 over ROI for the different CS computed for live interaction (left) and faked interaction (right) of subject number 7, both in the role respondent.....</i>	<i>129</i>
<i>Figure 70 : The spatial distributions of saccadic eye movements in talking, listening and thinking modes (from Lee, Badler et al. (2002)). The circle is supposed to be centred on the interlocutor.</i>	<i>131</i>

- Figure 71: Comparing original gaze patterns (top) with patterns generated using the statistical model driven by the sequence of cognitive states (bottom). Left: fixations labeled with ROI; right: fixations labeled with cognitive state. 132*
- Figure 72 : Our virtual talking face is driven by 12 facial degrees-of-freedom (see Gérard Bailly, Elisei et al. (2006a)). The eyes and eyelids movements are controlled by 5 degrees-of-freedom that captures the correlations between gaze and eyelids deformations (Gérard Bailly, Elisei, Raidt, Casari & Picot (2006b)). 132*
- Figure 73 : Entrance and transition probabilities. For each combination of CS and role a table of entrance and transition probabilities is calculated. The upper bar of each figure displays the probability at which the first fixation of an instance of a CS is directed to a given ROI. In the array below, the transition probabilities between ROI are displayed. The columns represent the previous, the rows the subsequent fixation target. At the cross points, the probability of transition from one to another target is indicated, coded as a gray scale level, whereas the darker the color, the higher is the probability. To indicate the reliability of these values, below the diagram, the number *n* of instances from which these values are calculated is given. A small number of samples results in less reliable probabilities. The diagram covers the fixation targets face, right, eye, left eye, mouth, else as well as no fixation detected and no transition, which are indicated as abbreviations on the margins. 134*

GLOSSARY

- accommodation: adjusting the curvature of the crystalline lens in order to focus the looked at scene on the retina, depending on its distance
- avatar: graphical representations of a virtual character that is at least partly controlled by a human agent
- behavioral implicit communication - BIC: behavior performed in the awareness of being observed and with the acceptance or even the intention to be observed and interpreted by others, but without this being the primary purpose of the behavior
- ClearView®: graphical interface for the configuration and control of the Tobii® 1750 eye tracker, as well as for the measurement and evaluation of data
- cone: photosensitive cell, does not saturate, appropriate for daylight view. Three different types of cones of different curves of spectral sensitivity: Cones of type 'S' have the highest sensitivity at a wavelength of 420nm (violet), cones of type 'M' at 530nm (green) and cones of type 'L' have a peak of spectral sensitivity at 565 nm (yellow). There are approximately 6.5 million cones in the retina.
- conversational agent: software agent, designed for linguistic communication typically between a machine and a human user.
- covert shift of attention: shift of attention that is not accompanied by a gaze shift
- crystalline lens: biconvex structure in the eye of a diameter of 9 to 10mm
- embodied conversational agent – ECA: conversational agent disposing of a bodily representation
- eye tracker: equipment to measure gaze direction
- fovea: small dip in the retina of a diameter of about 2mm, directly opposite the lens; contains the highest concentration of cones and produces the highest acuity of image resolution.
- initiator: person leading a dyadic conversational interaction, delivers the topic and content and directs its course (see also role and respondent)
- overt shift of attention: shift of attention that is accompanied by a gaze shift
- respondent: person taking a more reactive role in a dyadic conversational interaction, with the tendency to receive information and to respond to questions and proposals
- retina: inner neurosensorial layer of the eye, contains photosensitive cells
- rheme: part of an utterance that presents a new or specific contribution to the current context of a conversation (see also theme)
- rod: photosensitive cell, extremely sensitive to low intensities of light, not sensitive to color; there are approximately 130 million rods in the retina.
- role: we considered that there are two basic attitudes a person can adopt in interaction independent of turn taking – initiator and respondent

-
- semantically unpredictable sentence: grammatically correct sentence, but without any reasonable meaning; the individual words cannot be reconstructed using top-down semantic information
 - stimulus onset asynchrony – SOA: time delay between the onset of a stimulus and a cue demanding a reaction of a subject; typically used in experiments of shifts of attention in relation to visual stimuli.
 - talking head: 3D animation of a humanoid head able to produce visual speech articulation
 - TETserver®: software tool for direct online access to gaze data measured with a Tobii® 1750 eye tracker
 - theme: part of an utterance that links it to the present discourse and specifies what the utterance is about
 - theory of mind: ability to conceive the motivations, goals and strategies of another person
 - visual angle: angle produced by the maximal linear displacement of the looker's iris that cannot yet be perceived by the observer, calculated from the magnitude of the displacement of the iris and the distance between looker and observer; used as a measure of visual acuity

INTRODUCTION

The subject of this thesis is part of the effort of our laboratory to analyze and model the protocol of exchange of linguistic and paralinguistic information in human conversation. Special focus is put here on multimodal signals encoding this information, the aim being to determine when, how and why this information should be processed.

A lot is known now about the structure of acoustic speech signals and notably its prosody. These signals help interlocutors regulate the flow of information, subdivide the audio stream into meaningful units and communicate attitudes and affect. Other multimodal signals, with a stronger orientation towards visible manifestation also contribute to the pacing of this information. The body, hand, head, face and eye movements constantly inform our interlocutors on the attitude and emotions related to what we are saying as well as to what they are telling us. These gestures are such intimate parts of our communication channels that we even produce them when the interlocutor is not directly present. In this case, we are not only addressing imaginary interlocutors. While these gestures are not part of the information that actually reaches the recipient, they contribute to reducing our own cognitive load. They are part of our highly trained sensori-motor behavior and these motor activities are part of our proprioceptive feedback and participate in the proper triggering of our normal cognitive abilities. We cannot just stop behave as a social species when our interlocutor is not physically present.

This work is dedicated to the study of eye-gaze and particularly to its role in attracting our interlocutors' attention on specific objects in our environment and in keeping their attention focused on what we are saying. In face-to-face interaction, gaze plays many other functions notably in conjunction with the structuring of dialog and argumentation. Gaze is also strongly involved in the regulation of speech flow and turn taking. It plays a crucial role for the establishment of mutual attention and grounded interaction. Deictic capabilities of gaze and mutual attention often combine. Most of our social activities involve collaborative tasks for which the environment is deeply involved in dyadic conversations.

For virtual characters, these functions are vital to generate a sense of presence and consciousness of the environment. While a robot may demonstrate presence with its mere bodily appearance and show awareness of the environment when manipulating objects or avoiding obstacles, this requires more effort for a virtual character displayed on a screen. Appropriate animation of gaze as well as correct reactions to the gaze behavior of a human interlocutor, are powerful and necessary means to generate this impression of presence and awareness. Of course, this requires a rich scene analysis as well as appropriate models of interaction. The analysis and modeling of gaze behavior involve perception-action loops operating at different scopes. While smooth pursuit of a moving object involves low-level coupling between retinal perception and the control of eye gaze, attention towards elements referenced in the discourse and present in the scene often require analysis, comprehension and synthesis of dialogic acts over time scales of the order of a few utterances.

The final goal of this work is the modeling of gaze for a virtual conversational agent. Our methodology of investigation is mainly experimental. We have studied and characterized real-time interactions between humans involved in dyadic conversations as well as human-agent interactions. Our objectives are both precision and relevance. Precision is obtained via precise monitoring of gaze during interactions using up-to-date eye tracking techniques. The thesis presents detailed analysis of the multimodal scores, including low-level signals such as gaze, speech, facial movements as well as high-level information concerning related cognitive activities of the interlocutors that motivate the monitored signals. Relevance is achieved by scripting the interaction scenarios in such a way as to control the situation and ease the

interpretation of the observed behaviors. This work is thus highly focused and does not claim to provide a global sketch of all possible strategies to implement gaze skills in artificial agents. It provides however very detailed analysis and modeling of gaze patterns observed during well-controlled task-oriented interaction situations.

The thesis is organized in four main sections:

- Section 1 - 'State of the art' - provides a state of the art overview of the research on gaze and visual perception that is of importance in the context of our experimental work as well as for general comprehension. We also provide a general overview in the domain of embodied conversational agents that are the basis of our research platform.

The experimental work of this thesis is separated into two blocks.

- Section 2 - 'Gaze direction and attention' - describes two sets of experiments analyzing the impact of deictic gestures of a talking head on user performance in a simple search and retrieval task on a computer screen. We tested to what extent our talking head is able to generate persuasive deictic gestures, and how a user can benefit from these gestures. The results of these experiments show that positive assistance provided by the talking head reduces the subjects' processing time in a search and retrieval task. Furthermore, it is able to reduce the number of objects inspected during the search process, which we interpret as a reduction of cognitive load. It also appears that subjects manage to ignore deceptive indications suggesting that the impact of the talking head on the direction of attention does not dominate the subjects' conscious search strategies. An interesting result is the improvement of performances when an imperative utterance enhances the gaze and head gestures. This accentuation of the temporal incident of information transmission seems to be an important factor to direct attention. The multimodal gesture, specifying the location via the visual gesture and the timing of its interpretation via the audio channel, has an important impact on performance.
- The second block of experimental work, described in section 3 - 'Mediated face-to-face interaction' - is a first approach to precisely investigate the gaze behavior of interlocutors in close face-to-face interaction. In the case of gaze behavior in human interaction, the most important work originates from the 1970s. At that time, eye tracking facilities were far less developed and were mainly based on human observation. The temporal resolution of the measurements as well as the precision of the measure of gaze direction were rather coarse. Since then little has been gained on the knowledge about gaze behavior in face-to-face interaction. Our work on mediated face-to-face interaction is an attempt to renew research in this domain, with the perspective of enhancing gaze control of virtual characters. It consists in developing and implementing an experimental platform and scenario, and a series of experiments based on restricted task-oriented dyadic interactions. Based on the analysis of the experimental data, we developed a model for the animation of the gaze of an embodied conversational agent. The results confirm that the eyes and mouth are salient targets of fixations as already put forward in the works of other researchers. We show that gaze behavior varies in relation with the different states of the conversation. As expected, the gaze behavior varies between individual subjects. Over the interactions with different interlocutors, our target subject however showed a rather consistent and balanced repartition of gaze over the eyes and the mouth. In certain cases, there is a very obvious relationship between the segments of the interaction and observed gaze patterns. Our target subject tended for instance to look at the mouth while listening. This intensifies during moments where she expects new information compared to moments where she is listening to already known utterances. The eyes are however salient targets throughout the experiments and are fixated regularly. A relationship between gaze behavior and the occurrence of eye contact could not be found in our data.

It evidences neither avoidance nor search for eye contact. When our target subject is replaced by a prerecorded video, in order to put subjects into a faked interaction, this changes the subjects' gaze behavior in certain cases. This seems to vary with the extent to which the subject is convinced to be still interacting with a real person. The analysis of blink also puts forward a relationship between frequency of occurrence and segments of the conversation. The observations favored the hypothesis that increased attention leads to reduced frequency of blink. Furthermore, there is a remarkable tendency to blink when preparing to speak, especially when concurring with a larger head movement. This may be a protective reflex or a signal in the context of turn taking, similar to the aversion of gaze at the beginning of a turn.

- Section 4 - 'Conclusions' - summarizes the results we obtained and points out perspectives for future work on the topics raised.

1 STATE OF THE ART

Gaze is the main cognitive feature under study in this thesis. To lay a foundation for the considerations and discussions to follow, we will summarize in this chapter the fundamentals of human visual perception and its implication in human communication and interaction. We first explain the physiology of the eye and the functionality of its locomotive system. The characteristics of the eye are important to understand its visual appearance and its role in visual perception. The techniques used for the measurement of gaze direction are further described in order to give technical background for appreciating the cited reports and experimental work. We then discuss in detail how the outstanding prominence of eyes and gaze orientation as visual stimuli arises from the importance of the visual sense in human perception. We further outline what are the main functions of gaze in human communication and interaction.

1.1 THE PHYSIOLOGY OF THE EYE

In this section, we describe the configuration of the eye and the ocular locomotive system. Given the characteristics of eye movements, the related visual processing and the context in which these movements appear, different types of movement are distinguished. In this description, we incorporate blink as an activity closely related to the functions of the eye.

1.1.1 The anatomy of the eye

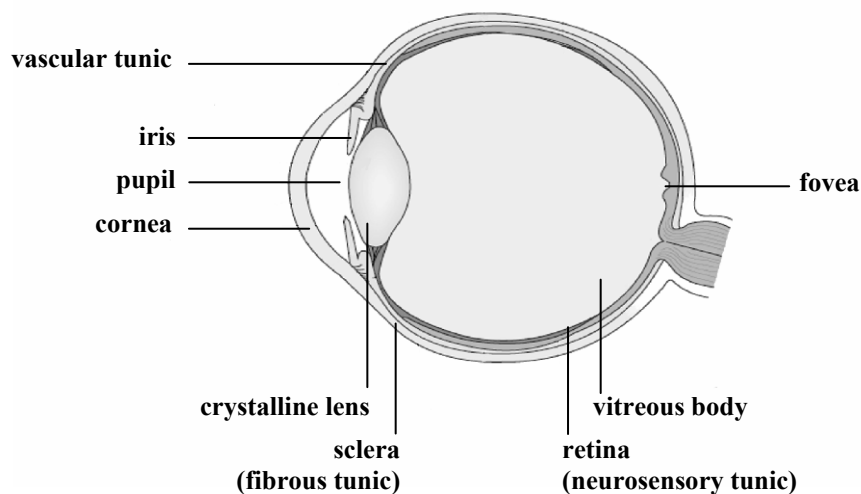


Figure 1: Anatomy of the eye

The basic structure of the eye consists of three main layers. They surround the 'vitreous body' of the eye, which is transparent and consists mainly of water. The outer layer of the eyeball is denominated 'fibrous tunic'. It maintains the shape of the eye and protects its inner parts. Around the 'iris', where the fibrous tunic is white, it is called 'sclera'. The front part that is translucent and covers the pupil is the 'cornea' (Encyclopædia Universalis <http://www.universalis.fr>).

The middle layer is denominated 'vascular tunic'. It contains the blood vessels that supply the retinal cells. The visible part of the vascular tunic is the 'iris' that contracts or dilates to change the size of its aperture for the regulation of the amount of light that penetrates through

the pupil and the 'crystalline lens'. The 'crystalline lens' is a biconvex structure of a diameter of 9 to 10mm. For adjusting the curvature in order to focus the image on the retina, the lens is deformed by surrounding muscles according to the distance of a gazed at object. This process is called 'accommodation'.

The retina is the inner 'neurosensory tunic'. It contains photosensitive cells of different types. 'Rods' are cells that are extremely sensitive to low intensities of light but cannot detect color. They typically saturate at an illumination corresponding to a dimly lit room. Their peak sensitivity is at a wavelength of 500 nm, corresponding to the color cyan. 'Cones' are photosensitive cells that do not saturate and are therefore more appropriate for daylight vision. There are three types of cones that have different curves of spectral sensitivity. Cones of type 'S' have the highest sensitivity at a wavelength of 420nm (violet), the cones of type 'M' at 530nm (green) and cones of type 'L' have a peak of spectral sensitivity at 565 nm (yellow). In total, there are approximately 130 million rods and 6.5 million cones in the retina.

The 'fovea' is a small dip in the retina of a diameter of about 2mm, directly opposite the lens, which contains a high concentration of cones. At the centre of the fovea, in an area of 0.4 mm of diameter, is the highest concentration of cones that are exclusively of the type L and M. This area corresponds to 1.4° of visual angle. Rods do only appear outside of the fovea, increasing in density with increasing distance from the fovea until their concentration becomes superior to the concentration of cones in the peripheral parts of the retina (see Figure 2). This strongly heterogenic configuration of the retina results in a different functionality of the respective parts. High acuity foveal vision is limited to a few degrees of visual angle. The periphery of the retina is highly sensitive even to very low illumination, but is inadequate for the perception of color.

For the orientation of gaze, the movement of the eye in the orbit is controlled by six muscles (see Figure 3), the 'medial rectus' (mr), 'lateral rectus' (lr), the 'superior rectus' (sr) and 'inferior rectus' (ir) as well as the 'superior oblique' (so) and 'inferior oblique' (io). The medial and lateral rectus muscles generate the horizontal components of eye rotations. Vertical and torsion movements are accomplished by the activation of combinations of the pairs of superior and inferior rectus and the superior and inferior oblique muscles (Sparks (2002)). These muscles are innervated by motor neurons in the cranial nerve nuclei. The motor neurons generate a burst of spikes over a period of time that is approximately equal to the duration of a rapid eye movement (saccade). During intervals in which the eyeball is relatively stable (fixation), the motor neurons discharge at a constant rate that is a linear function of the position of the eye in the orbit.

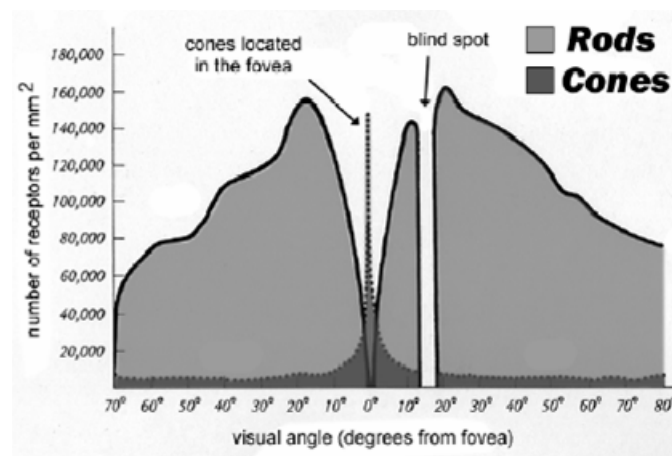


Figure 2: Diagram of the concentration of rods and cones on the retina around the fovea.



Figure 3: Top and front views of the right eye and the extra-ocular muscles: medial rectus (mr), lateral rectus (lr), superior rectus (sr), inferior rectus (sr), superior oblique (so), inferior oblique (io); Sparks (2002).

1.1.2 Eye movements and blinks

The movement of the eye is categorized in different types with different characteristics. The different types of movement are partly controlled by different neural systems but are combined via a final common path which leads to the six muscles controlling the eyeball (Robinson (1968) cited in Argyle & Cook (1976)).

1.1.2.1 Saccades

In general, humans suppose to have a clear and detailed visual perception of the environment. In reality, we only perceive a very small sector of a visual scene at high acuity that corresponds to 2° or 3° of visual angle (Liversedge & Findlay (2000)). This is the section of the scene that is projected onto the fovea and the parafoveal regions. The eye scans the scene successively, executing fast movements to direct the gaze at different areas. Their projection falls onto the fovea where it is perceived at maximum resolution. We integrate the respective parts of the visual scene that we see at high resolution into the perception of the whole scene generating the impression of a clear vision of the entire environment. The high velocity eye movements executed to orientate foveal view are denominated 'saccades'.

For the generation or programming of saccades, two independent cognitive processes in the brainstem control the location of the saccade target and the start of its execution. The horizontal component of a saccade originates from premotor neurons in the pons and medulla. The vertical component is produced by premotor neurons in the rostral midbrain. There is a relation between the amplitude, duration and velocity of the saccadic eye movement and the number of spikes, burst duration and peak discharge rate of such neurons. The horizontal and vertical components of saccades are coordinated in such a way that they results in a straight, not curved movement.

Voluntarily executed saccades that are provoked by inner mental processes are denominated as endogenous or top down generated. They are preceded by a shift of attention. Exogenous, stimulus driven saccades are triggered by external events such as visual or audio stimuli and considered as bottom up generated (Godijn & Theeuwes (2003)). The delay between the onset of a stimulus that attracts attention and the onset of a saccade towards this stimulus lies typically within 100 to 300 ms.

Between two consecutive saccades, there is a minimum interval of a refractory period of a duration of 50 to 200ms. During saccades, visual perception is very restrained. Once started, a saccade is a ballistic movement that cannot be interrupted or its destination be changed. The selection of a target is therefore made before the start of the saccade.

The duration and velocity of a saccade depend on its magnitude (Freedman & Sparks (1997)). There is a linear relationship between the saccade amplitude and its duration over a wide range of magnitudes. The duration is in the range of 100ms and the velocity can be of several

100 degrees per second (400 – 800). Saccadic eye movements are often accompanied by a head rotation in the same direction. Under natural unrestrained conditions, saccades have an average magnitude of 15 to 20 degrees and large gaze shifts usually include a head rotation.

1.1.2.2 Gaze shifts as a combined movement of eyes and head

The coordination of eye and head movement during gaze shifts of different amplitude has been investigated by Freedman & Sparks (1997) in experiments with rhesus monkeys. In experiments where the heads were restrained, they measured smooth eye velocity profiles during horizontal gaze shifts (see Figure 4). The peak velocity of these profiles is function of the saccade amplitude, which saturates at a velocity of about 1100 degrees per second, as shown in Figure 5 (left).

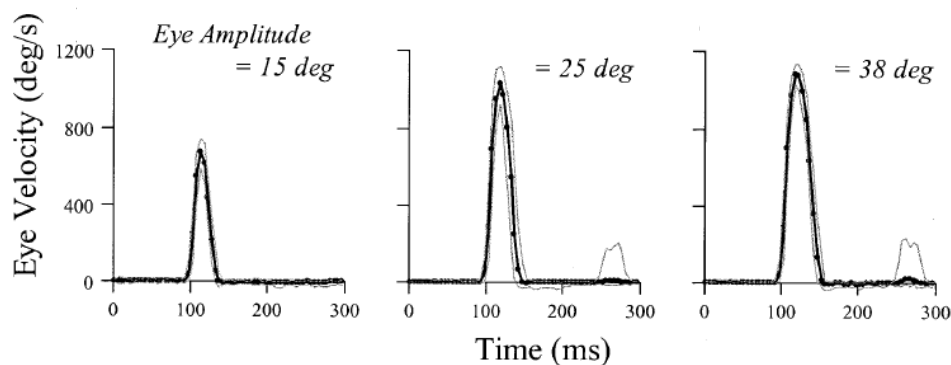


Figure 4: Velocity curves observed during horizontal gaze shifts of different amplitudes, in experiments on rhesus monkeys with heads restrained (Freedman & Sparks (1997)).

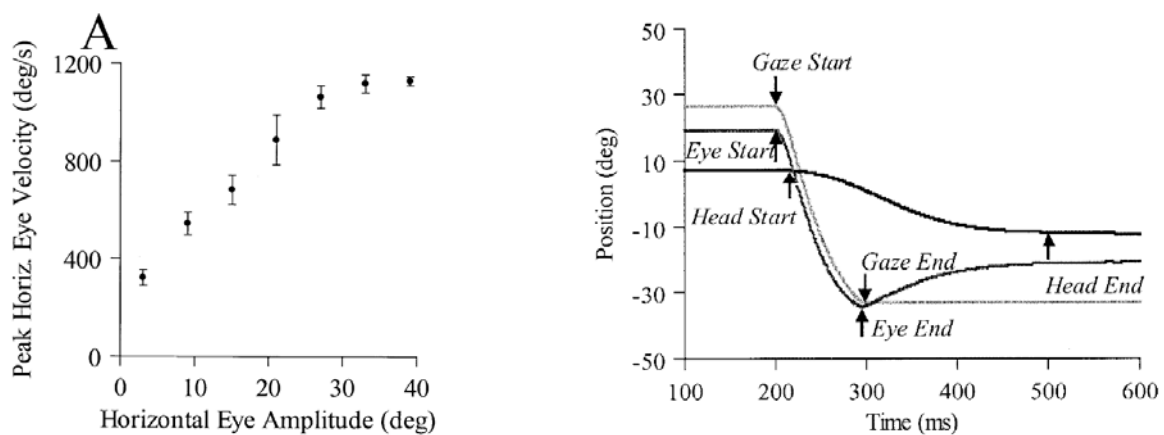


Figure 5: Left: Peak velocities as observed for horizontal eye movements of different amplitudes. Right: Schematic course of head and eye position of a combined movement of gaze shift (Freedman & Sparks (1997)).

The course of eye movement becomes more complicated when the head is unrestrained, especially when the pupil is not at the centre position of the eye at the beginning of the gaze shift. Figure 5 (right) shows a schematic diagram of the head and eye position during a fast gaze shift. The eye overshoots the final position relative to the head in order to direct the gaze toward the target as fast as possible. Whereas the gaze is already directed to the target, the head continues its movement until it has attained its final position. At the same time the eyes performs a compensatory movement with the same velocity as the head but into the opposite direction in order to maintain gaze stable until the head has also attained its final position. Ramat, Schmid & Zambardi (2003) proposed a model that accounts for the coupling of head velocity and gaze control.

1.1.2.3 Smooth pursuit

When following a moving object, the eyes perform a continuous movement, denominated 'smooth pursuit'. Smooth pursuit movements are synchronized to the movement of the observed object. Such movements cannot be executed volitionally. With an angular speed of 30° to 90° per second during smooth pursuit the eyes move much slower than during saccades (Barattelli, Sichelschmidt & Rickheit (1998)).

1.1.2.4 Fixations

The most important periods for visual perception are fixations that are periods of relative stability between saccades. During fixations, the eyes are not completely stable and still. Movements of different origins change the position or configuration of the eyes. The small movements occurring during fixations are essential for visual perception, since the optic sensors on the retina are sensitive to changes in light intensity. A perfectly stable projection of a scene onto the retina would make visual perception vanish. A constant image projection on the retina, generated for instance with special contact lenses fixed to the eye ball, disappears very quickly from view (Pritchard, Heron & Hebb (1960) cited in Argyle & Cook (1976)).

Movements that occur during fixations are for instance microsaccades that correct small drift movements of the eyes during fixations. They are accompanied by a high-frequency tremor of the eye. The orientation of microsaccades is mainly horizontal or vertical, depending on whether they are monocular or binocular (see Figure 6). The vestibulo-ocular reflex is responsible for eye movements that compensate for head movements during fixations, in order to keep the projection of the gaze target in the centre of the retina.

To adapt to the distance between the eye and a fixated object, the eye executes vergence movements to direct both eyes towards the same target. Accommodation of the eye deforms the optical lens according to the distance in order to obtain a focused image on the retina. Contraction or dilation movements of the pupil are executed to regulate the amount of light falling on the retina depending on the brightness of a scene. The pupils may also dilate as reaction to increased arousal, mainly when it is of positive nature.

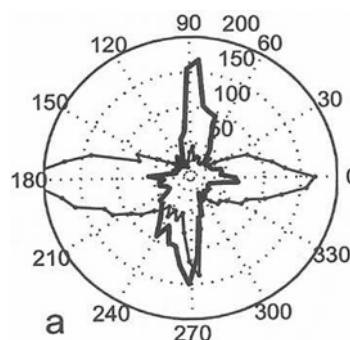


Figure 6: Angular distribution of microsaccades: thin line – binocular; bold line – monocular (left or right eye) (from Engbert & Kliegl (2003))

1.1.2.5 Blinks

A blink is the rapid movement of the eyelid, closing and reopening, with the effect of moistening and cleaning the surface of the eye or as a protection reflex. Argyle & Cook (1976) report an average rate of 12.5 blinks per minute for adults. Itti, Dhavale & Pighin (2003) cite authors reporting similar blink rates and in relation with augmented attention observations of increased blink rate.

Ponder & Kennedy (1927) examined blink frequency under different conditions. In a population of 50 persons, they found very different blink behaviors during normal reading. Distributions of inter-blink periods were very diverse and highly depending on each

individual. For constant experimental conditions, the distributions observed with the respective individual were however very constant over time, even when checked at different instances over long periods of time. Most of the subjects tended to have short inter-blink intervals with a distribution showing decreasing amount of blink for increasing inter-blink durations. For these subjects, there was however a strong variation of mean duration and constancy of frequency varied. For other subjects they observed bimodal distributions, completely irregular distributions or distributions that are symmetrical around an elevated mean value.

Searching for the origin of the impulse triggering blinks, they concluded that the normal and periodic movements of blinking are not dependent on afferent impulses arising from the retina, the cornea, the conjunctiva, or extrinsic muscles and that they are not necessarily dependent on impulses passing in the second, third, fourth, fifth, or sixth cranial nerves. Major changes of air humidity or illumination and anesthesia of the cornea and conjunctiva had no persistent effect on blink rate. After such changes of experimental conditions, only transient effects - if any - could be observed. The fifth cranial nerve transmits impulses from superficial parts of the eye such as eyelashes or eyelids. Subjects with a defective fifth cranial nerve due to a surgery did still show normal blink rate. Even blind people show normal blink behavior. Hence, Ponder & Kennedy excluded the possibility that the optical nerve is required for a normal blink rate. They found no evidence that the movements are reflex at all, as they could not identify an afferent path, the destruction of which would cause blinking to cease. They assumed the movements to be of central origin and caused by a periodic and more or less regular discharge of impulses through the seventh nerve. From the fact that blinking completely ceases associated with a localized lesion of the brain, which is found in post-encephalitic Parkinsonism, they concluded that the basal nuclei region, especially the caudate and lentiform, is closely associated with the generation of normal blinking.

There is however clear evidence that blink rate can be influenced by different factors such as irritation of the conjunctiva or stimulation of the auditory nerve. Ponder & Kennedy report also differences of mental condition such as excitement or tension that produce increased blink rates (also reported by Harris, Thackray & Schoenberger (1966) cited in Argyle & Cook (1976)). They assumed that this was a kind of relief mechanism used to release nervous energy similar to agitated hand movements that can be observed in the same situations.

1.2 MEASUREMENT OF GAZE DIRECTION

For the measurement of gaze direction, different methods have been developed. As technical possibilities evolved, the methods used in scientific experiments changed a lot over time. The experimental conditions, such as the setup and task, but especially the applied method of gaze measurement are important criterions for the evaluation of acquired data and results. We therefore describe the most important techniques that have been used in the investigations discussed in the following.

1.2.1 Human observers

The easiest method to monitor gaze direction is to use human observers. This has been the most common technique in the early stages of research on gaze. Argyle & Cook (1976) report that most of their experiments and the work they cite have been performed using human observers in various ways. Common parameters observed to characterize gaze, are mutual gaze, total gaze, frequency and length of glances, often with the additional distinction between gaze while listening and gaze while speaking. To take down the observations of gaze, pen recorder devices that an observer manipulates with keys, stopwatches or simple notes on paper have been used. The observers were either directly interacting with the

subjects or placed more or less hidden from the interacting subjects. In some cases, considerable effort was made to conceal the fact that the subjects were observed at all. One-way screens or hidden holes in a wall have been used for this purpose for instance. Sometimes a continuously staring confederate was used to simplify the measurement of mutual gaze. Gazes of the subject towards this confederate were automatically considered as instances of mutual gaze.

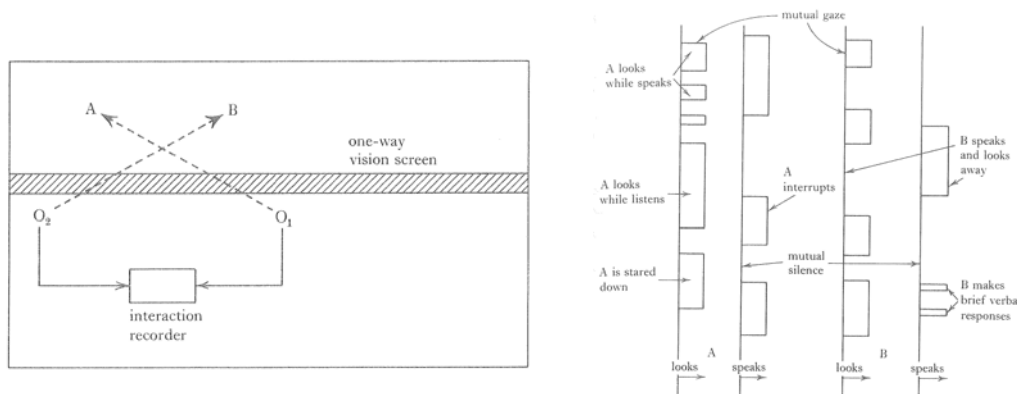


Figure 7: Left: Schematic drawing of an experimental arrangement for the gaze monitoring of two subjects (A and B) by two observers (O₁ and O₂) behind a one-way screen. Right: Pen writer protocol of the interaction of two subjects listing intervals of gaze and speech (Argyle & Cook (1976)).

Kendon (1967) used cameras to take photographs of the subjects at a constant frequency of 2 Hz for later evaluation. He recorded the audio signals of the conversation along with synchronization signals sent by the camera at every shoot. To assure synchrony between the recordings of two subjects he photographed both with a single camera using a mirror for the image of one of the subjects.

The techniques relying on human observers are subject to shortcomings due to human fallibility such as for instance the leak of temporal precision due to reaction time. The breakdown of gaze shifts into saccades with durations of only a few hundred milliseconds is impossible with these techniques. Furthermore, the resolution of gaze direction is very coarse and cannot reliably be associated with an exact target, as the location of the stimulus cannot be monitored precisely. Relative gaze shifts may have to be associated with one of many possible targets at close local proximity. Argyle & Cook mention different studies by other authors that evaluate the agreement between observers as well as their self-consistency of gaze measurements. These report different but in general high values of reliability. This is however no guaranty of validity of data. It may still be false as the consistency may result from similar strategies.

1.2.2 Mechanical techniques

The first precise mechanical technique was used by Yarbus (1967). He used a contact lens with strong adhesion that therefore follows exactly the movements of the eye. He attached a mirror to the contact lens that reflects the light to expose a photosensitive layer. This technique is very precise and even allows projection of the gaze recordings onto the stimulus when images are used as such. Temporal information of the gaze trajectory however is lost. Another technique mentioned by Argyle & Cook, is the use of a mechanical lever or a magnetic coil attached to the contact lens to measure the gaze direction. The latter may even be appropriate for use during relatively unrestricted dialogue. However, the application of such mechanical devices is very uncomfortable for the subject and interferes with blinking.

1.2.3 EMG techniques

Jacob (1993) describes an electronic eye tracking technique using electrodes placed on the skin around the eye to measure changes in the orientation of the potential difference that exists between the cornea and the retina. This only allows measurement of relative eye movements if the head movements are not restricted and provides only low accuracy.

1.2.4 Video techniques

Nowadays, mainly video-based techniques are used exploiting visible features of the eyes or reflections of light on the eye. The outline of the pupil as boundary between the sclera and iris is such a feature. To generate reliable reflections on the cornea usually infrared light sources are used. Computing two features per eye, gaze orientation can be estimated independently of head movement. Head mounted eye tracking devices may be equipped with an additional scene camera for identifying the targets of the recorded gaze directions. Some devices can also monitor the head orientation. The orientation of the head and the gaze orientation relative to the head orientation can be integrated to determine the absolute direction of gaze in space. Eye trackers as desk mounted units or integrated into a computer screen allow relatively unrestricted head movement of the subjects without disturbing their visual appearance. High sample frequencies up to several 100Hz and high accuracy in measured angles can be achieved with appropriate cameras. The video recordings used for the calculation of gaze orientation can also be exploited to determine other parameters of the eye such as pupil size and blinking.

1.3 VISUAL PERCEPTION OF EYES AND GAZE AS SPECIAL STIMULI

Eyes and gaze are special stimuli for human perception in every culture. Even in animals, they have been reported to trigger special reactions and behavior. Obviously being looked at by a predator may precede an attack and an immediate reaction to such a gaze is a clear advantage. This may explain why for animals direct gaze is often interpreted as a signal of threat or aggression. Argyle & Cook (1976) report that animals often react to eye like stimuli and that eye spots on animals such as butterflies, birds, fish and snakes may keep off predators. Some animals avoid gaze contact the more, the closer they are to each other, which may be necessary to reduce arousal. Argyle & Cook further report that gaze also plays an affiliative function only in primates and man. However, with humans it is of course also observed in the interplay of dominance and hierarchy. They summarized that for lower animals, eyes and eyespots serve as a stimulus to produce avoidance of predators. For animals higher in the evolutionary scale, gaze produces arousal and gaze direction may be decoded to deduce the location or object that another animal's attention is directed to. Human infants develop quickly a sophisticated processing of eye gaze until it adopts an important function in social interaction between adults.

In the following, we discuss in detail how the sensitivity of visual perception to eyes and eye like stimuli develops in humans, how precise the estimation of the direction of another's gaze is, and how it is interpreted as a deictic cue and may trigger shifts of attention.

1.3.1 Visual perception of eyes and gaze by children

The research on gaze behavior of children and their sensitivity to gaze of others is of great interest for the understanding of the outstanding meaning of gaze in human interaction. It helps to clarify whether the special cognitive processing of eye and gaze are innate in humans. Early studies on gaze perception by children are reviewed by Argyle & Cook. They led them to assume that the importance of eyes compared to mouth or nose must be innate. Ahrens

(1954), Spitz & Wolf (1946) (cited in Argyle & Cook (1976)) report that the view of faces, necessary to produce reactions of an infant, develops over its age (see Figure 8). At the age of about two months, a pair of eyes is sufficient to produce smiling, whereas a face with the eyes covered or a profile with only one eye visible will not. At each age there seems to be an optimum level of complexity for visual stimuli that increases with age. Klaus, Jerauld, Kreger, McAlpine, Steffa & Kennell (1972) (cited in Argyle & Cook (1976)) found that already at the first hour after birth infants turn their head towards visual and auditory stimuli. Argyle & Cook suppose that in this context the fact that the distance of 20cm, to which the accommodation of newborns is fixed, may favor the interest of infants in faces and eyes. This distance corresponds roughly to the distance of the mother's face during feeding.

Argyle & Cook cite several further studies that give evidence that infants develop an increasing interest for eyes and faces or similar stimuli whereas at the beginning the pattern of contrast of eye like stimuli seems to be sufficient. They report the preference of actual face images and the distinction of orientation of faces to develop during the first 3 month of age (Fantz (1965), Bond (1972) cited in Argyle & Cook (1976)). At the beginning, this preference is observed as changes in smiling, looking and heart rate. Later the infants are observed to follow gaze shifts and try to locate the target of another person's gaze. The progress in reaction to gaze is often observed to be rewarded by the caregiver by increased attention and playing. This probably favors and augments the reaction of the child to gaze. In fact, such playing often involves mutual gaze such as in playing 'peek-a-boo' (Bruner & Sherwood (1976) cited in Argyle & Cook (1976)).

L.A. Symons, Hains & Muir (1998) confirm the eye region as a preferred gaze target for very young infants. They observed direct gaze of adults towards children to result in longer gaze of the children and more smiling. In an experiment, they assessed infant sensitivity to a small shift of an interactor's eye gaze (approximately 5 degrees at a distance of 65cm). They found that in the case of horizontal deviations, infants of about 5 month of age are sensitive to these very small shifts in an interactor's gaze and smile more for an adult who maintained eye contact. For vertical deviations however, this effect could not be observed. Symons et al. interpret the sensitivity of young infants to small deviations in an interactor's eye gaze as evidence that eye-gaze information is a crucial cue used by infants in the development of their social interactions.

A recent study on the perception of gaze by very young infants was conducted by Farroni, Csibra, Simion & Johnson (2002). Their work gives important information to clarify whether the special sensitivity of humans to eye gaze is innate or acquired through experience. They hypothesize that the detection of preferential attention in human newborns to perceived faces with direct gaze provides strong evidence for an innate sensitivity. In two experiments, they demonstrated that from birth there is a special sensitivity to direct eye contact.

The first experiment tested the ability of 2- to 5-day-old newborns to discriminate between direct and averted gaze. With a light signal, the infant's attention was attracted to the center of a screen and once they fixated the centre, an image with direct gaze was shown on one half of the screen and an image with strongly averted gaze was shown on the other half of the screen. They found that the infants had a significant tendency to look longer and more often at the image with direct gaze.

In the second experiment, they measured the brain activity of 4-month-old infants, to assess neural processing of faces when accompanied by direct eye gaze in contrast to cases where gaze is averted. The transient brain activation has been monitored measuring ERPs (event-related potentials). In this experiment, images with either direct or averted gaze were presented to the subjects. For direct gaze a significantly stronger effect was observed (more negative amplitude in putative ERP component 'infant N170'). The experimenters consider the 'Eye Direction Detector (EDD)' that detects the presence of eyes (Baron-Cohen (1995)),

their direction and behavior, as possible explanation of this effect (see also section 1.4.2, page 45). However, they favor the hypothesis proposed by Johnson & Morton (1991) (cited in Farroni *et al.* (2002)) that there are subcortical circuits that support a primitive representation of high-contrast elements relating to the location of the eyes and mouth, to which a face with direct gaze corresponds better than one with averted gaze.

Summarizing their results, Farroni *et al.* conclude that from birth, human infants prefer to look at faces that engage them in mutual gaze and that, from an early age, healthy babies show enhanced neural processing of direct gaze. They consider that the contrast between direct and averted gaze results in a rather small psychophysical difference. They suppose therefore that the differences observed at birth are due to a fast and approximate analysis of visual input, dedicated to find socially relevant stimuli for further processing, rather than the mere preference of certain spatial frequencies. They consider the very early sensitivity to mutual gaze as the major foundation for the later development of social skills.

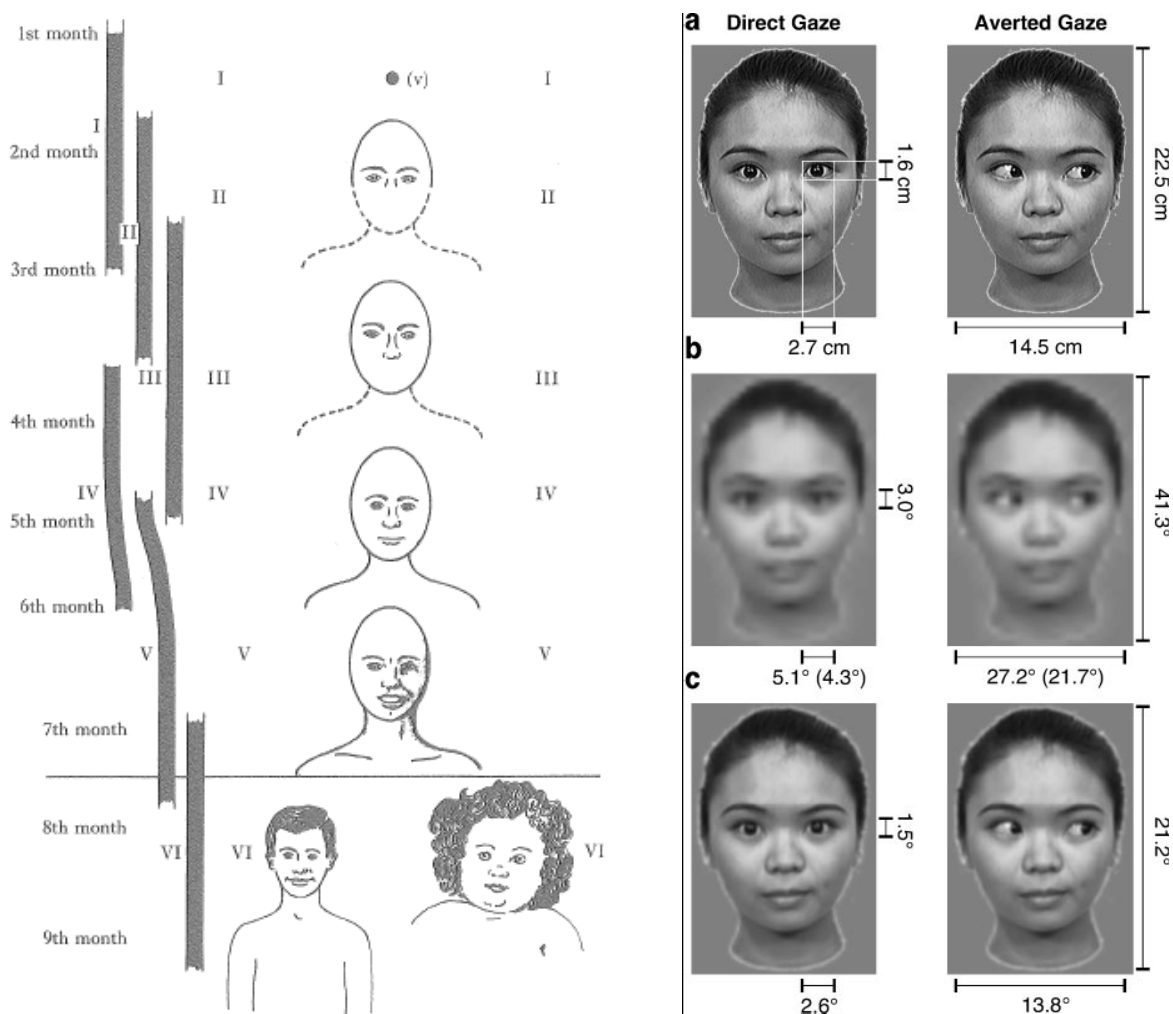


Figure 8: Perception of faces and face-like stimuli by children. Left: Overview over the type of stimulus needed to produce increased smiling of infants and its development over the first 9 months of life, as observed by Ahrens (1954). Right: Experimental stimuli used by Farroni *et al.* (2002), along with information about viewing angles of special regions, when fixated or when in the periphery (in brackets). (a) Direct or averted gaze (b) Images with a resolution according to the average visual acuity of newborns (c) Images with resolution according to average visual acuity of 4-month-old infants.

1.3.1.1 Gaze, speech and pointing

Another very interesting aspect in the development of social interaction of children is the development of pointing and its relation to the development of speech and visual perception.

Goldin-Meadow & Butcher (2003) claim that gestures and especially pointing gestures are highly related to the evolution of speech production in children. They consider gesture and speech to be integrated into one system of communication, both semantically and temporally. In children, they found the use of one-word utterances joint with pointing gestures to be a first step towards two word sentences and its onset to be predictable from the onset of the production of combinations of one-word sentences with gestures that convey different information.

Masataka (2003) reports that there is an increasing frequency of index finger extension until the age of 12 month. It then decreases and is accompanied by an increase of index finger pointing. Reaching and grasping gestures however, did not vary significantly over this time. Masataka concludes that this is a strong argument that pointing does not evolve from grasping gestures towards objects that are out of the reach of the infant.

1.3.1.2 Gaze and joint visual attention

Butterworth (2003) investigated pointing in the context of joint visual attention and observed several steps in the development of infants from following the gaze direction of another person towards pointing as fully developed communicational gesture. He observed that very young infants at the age of less than 9 month are able to follow another person's gaze shifts, which he interprets as a first form of joint visual attention. These children are however not able to localize target of the gaze shifts if several objects are in the visual field. Brooks & Meltzoff (2005) confirm these observations with their own experiments. They found that infants are beginning to follow movements of the head at the age of nine month. At about ten or eleven month, they are following head and eye movements whereas they are more likely to follow shifts in head orientation when the eyes are open.

Butterworth reports that at the age of about 10 month infants start to point at objects, followed by pointing at the other person as a kind of checking the attention of this person. He considers this as evidence for the communicative intend of the behavior. This is encouraged by the observation that infants only point when other persons are present. Their pointing behavior changes at an age of about 16 month, in the way that the infants first check the attention of another person before they point at an object. This reflects the understanding that mutual attention is necessary for communication. At this age, infants are also able to correctly localize fixated targets even when in the periphery of their visual field. This however is not the case for objects situated behind them. Their visual attention is still limited to their visual space. At the same age, infants start to interpret hand-pointing gestures. Butterworth considers the onset of pointing as an important step towards the understanding of the relation between objects and their names, as pointing connects visual referencing to a concurrent utterance. In the development of infants, pointing is important to establish the relation of identity between the two. This accentuates the importance of eyes and the interpretation of gaze in the context of joint visual attention. The capability to correctly interpret gaze direction and pointing gestures, is crucial for the development of language as well as for the establishment of a common space of interaction and grounded communication.

1.3.2 Perception of gaze direction

This chapter discusses the results from various studies on the perception of another person's gaze direction. A lot of work has been dedicated to the measurement of the precision with which another person's gaze direction can be estimated.

To be able to compare the measure of acuity of gaze estimation between different studies, independent of the experimental setup, we use the visual angle as proposed in Gibson & Pick (1963) and Cline (1967). The distance between looker and observer may for instance vary between setups. The visual angle is the angle produced by the maximal linear displacement of

the looker's iris that cannot be perceived by the observer. It is calculated from the magnitude of the displacement of the iris and the distance between looker and observer.

Some of the authors speculated about the features of the eye that are influencing the estimation of gaze direction. The work of Ricciardelli, Baylis & Driver (2000) is dedicated especially to this question. It is summarized at the end of this section.

As a general statement, the studies on estimation of gaze direction confirm eye-directed gaze to be a very special stimulus. Very important for our work are the results indicating that gaze direction of faces displayed on video screens can be perceived with comparable precision as physically real faces.

1.3.2.1 Precision of estimation of gaze direction

Gibson & Pick analyzed how a trained looker's gaze varying in the horizontal angle field is perceived by observers. They hypothesized that the perception of another person's gaze direction is not based on the appearance of the eyes alone but is estimated from the combination of estimates of head and eye orientation. They supposed asymmetries of the face and visible parts of the sclera around the iris to be of special importance in this context.

In their experiments, Gibson & Pick positioned a subject in front of a trained looker that fixated different targets on a line behind the observer, as shown in Figure 9 (left). The centre target coincided with the bridge of the observer's nose. While fixating the targets, the looker adopted one of three possible head postures (30° left, 0° , 30° right). Subjects were asked to estimate whether the looker directly fixated them. For the analysis they measured the number of positive judgments.

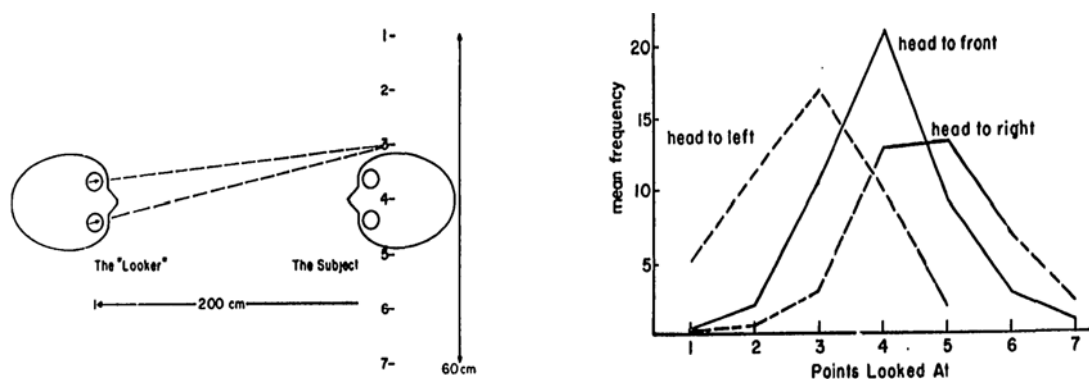


Figure 9: Left: Schematic view from above on the experimental setup used in Gibson & Pick (1963). Right: Results in Gibson & Pick (1963), measured as frequency of positive answers to the question whether the gaze of a looker was directed towards the subject or not, when the looker adopted different head orientations relative to the observer (30° left, 0° , 30° right).

As a measure of acuity of gaze perception, Gibson & Pick used the standard deviation resulting from the answers of all subjects grouped for head position. They found that for straight head orientation, the horizontal acuity is of $0.93'$ (minutes of arc) and of $1.08'$ for the turned head position. Furthermore, they observed a constant error of overestimated gaze angles for the turned head position. This means that gaze directions of the looker are perceived as eye-directed gaze, when they actually lie in an angle between the direction of head orientation and the direction of gaze that would actually aim at the eyes of the subjects.

The results by Gibson & Pick cannot unambiguously be interpreted, as in the report it is not always clearly explained whether the discussed directions refer to the looker's or the observer's point of view. The results of Cline can therefore not directly be compared to the results of Gibson & Pick. In Anstis, Mayhew & Morley (1969) this problem is mentioned, too. On the one hand Cline agrees with the principal findings of Gibson & Pick (1963) and

uses a similar expression to describe the quality of the observed constant error. On the other hand, he explains that the perceived direction is estimated from the interaction of head orientation and eye direction and would lie between the two. This however contradicts the explanations given by Gibson & Pick according to our understanding.

Compared to the experiments of Gibson & Pick, Cline used expanded viewing angles and additionally investigated the deflection in vertical direction. This experimental setup was enhanced by a half-silvered mirror (see Figure 10). This allowed presenting the gaze targets to the looker without exposing him to the view of the subject, which might be a source of distraction. The setup nevertheless gives the impression to the subjects to be in the plane fixated by the looker. Furthermore, the reactions of the subjects were monitored more detailed. Subjects were asked to mark the gaze target of the looker on a template offering 65 possible targets of which only thirteen corresponded to actual fixation targets of the looker. In addition, subjects had to report when they had the impression of being directly looked at.

Cline reports a measured horizontal acuity of $0.51'$ and a vertical acuity of $0.88'$ for the centre target. According to these measures, horizontal acuity is better than vertical. Both acuities significantly decrease for targets away from the centre of the face. Cline concludes that gaze directly directed towards the observer is perceived as a special stimulus. The finding that acuity decreases as targets move away from the centre position, contradicts the previous findings of Gibson & Pick.

Investigating the influence of the orientation of the head turned by 30° Cline found an increased constant error for the estimation of gaze direction and a decrease of acuity. When the looker fixated targets at different angles but with the head aligned with gaze direction the constant error was rather small, but acuity was still low for targets away from the centre.

Cline found no differences between the results obtained with the upper part of the looker's face in comparison to the full face view as stimuli. He concluded that only the eye region is exploited for the estimation of gaze direction and that it provides sufficient information. It must however be noted that in this part of the study the head was always directed straight ahead and only the orientation of eyes varied.

Anstis et al. extended the previous studies and hypothesized that the position of the pupil relative to the visible parts of the eye is the crucial information necessary to determine gaze direction. They realized two different experimental scenarios of face stimuli fixating targets at various visual angles. In one scenario, they used a person as stimulus, in the other a picture on a television screen (see Figure 11).

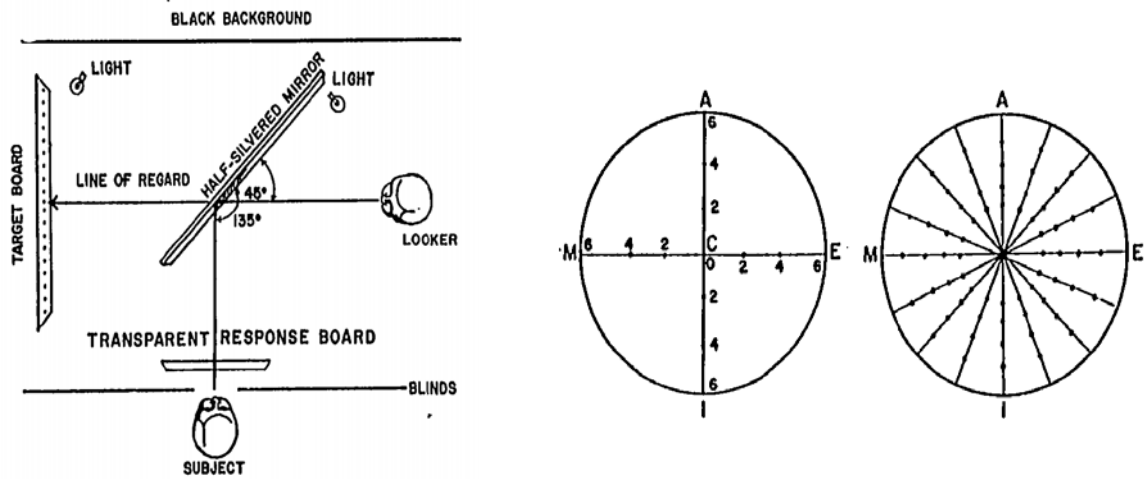


FIG. 1. PLAN OF APPARATUS

Figure 10: Left: Schematic representation of the experimental setup as used in Cline (1967). Right: Front view of target board containing the fixation targets for the looker (left) and response board on which subjects were asked to mark the estimated fixation target (right).

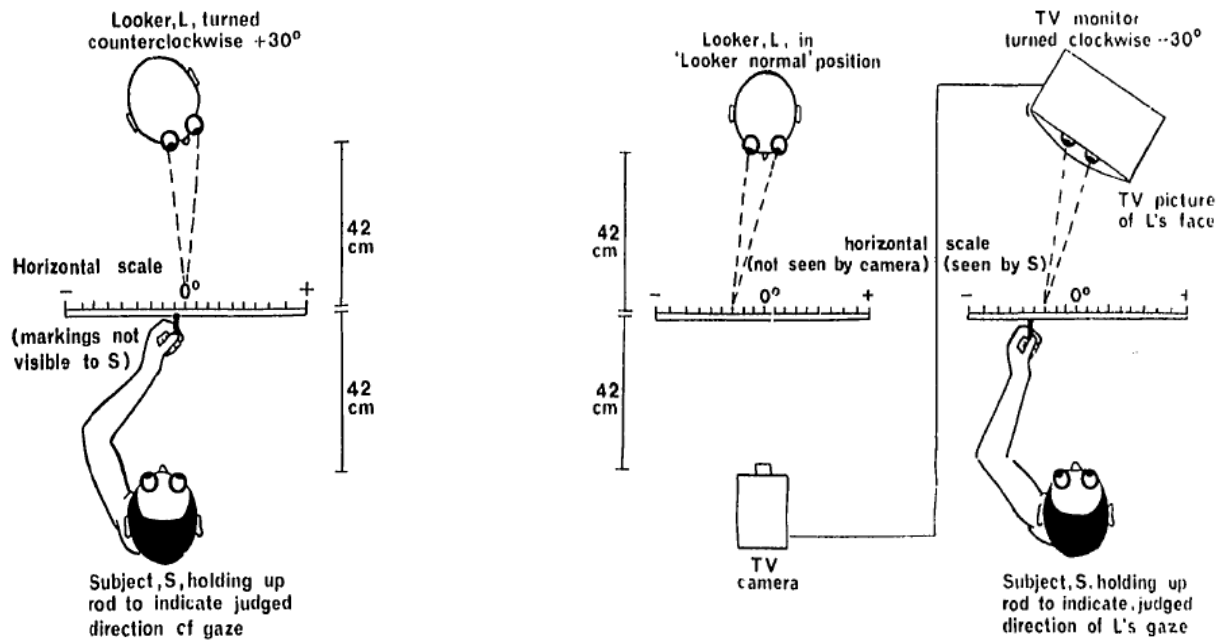


Figure 11: Experimental setups as used in Anstis *et al.* (1969). Left: Schematic view from above on setup with a looker fixating targets and a subject marking the estimated fixation target. Right: Similar setup but with a camera to acquire the stimulus that is presented on a TV screen turned at different angles relative to the subject.

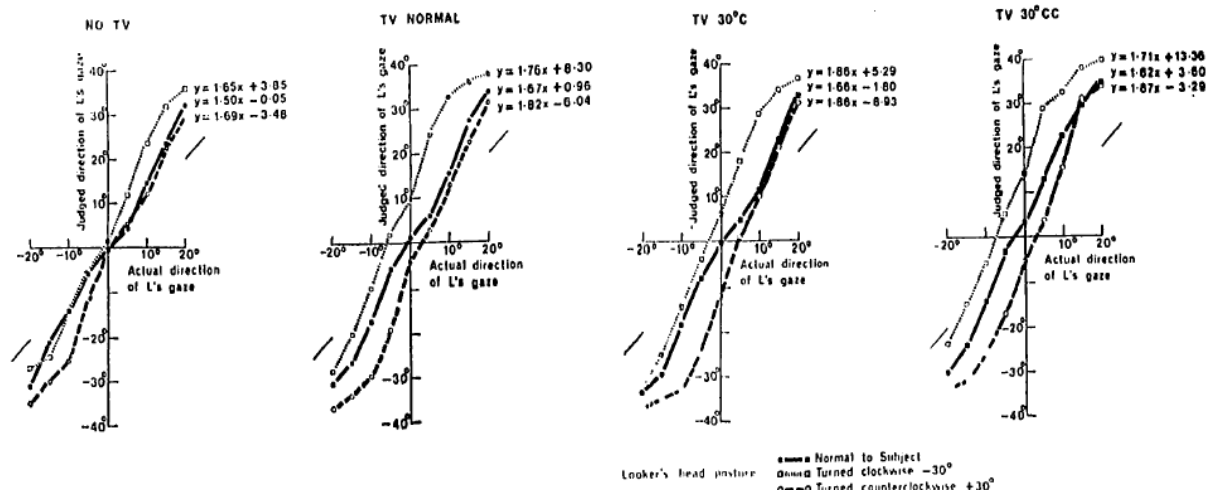


Figure 12: Charts putting the gaze direction as perceived by the subjects in relation to the actual gaze direction of the looker. The bold line represents straight head orientation of the looker, the thin line clockwise turn of the lookers head and the dashed line counter-clockwise turn of the lookers head.

Conditions from left to right: Direct presentation of looker; looker presented on TV screen positioned perpendicular to the subject; looker presented on TV screen turned clockwise; looker presented on TV screen turned counter-clockwise (Anstis *et al.* (1969)).

While the interpretation of the results reported by Gibson & Pick is ambiguous, the interpretation of the results in Anstis *et al.* is clarified in the text. They explain that the subjects had the imprecision of being looked into the eyes, when the lookers gaze was directed to the subject's ear while the head was orientated towards the subject's shoulder of the same side. This is a non-ambiguous statement of how to interpret the angles used to describe the results. To our understanding it correspond to the observations in Gibson & Pick, but contradicts the findings by Cline.

As a general result, Anstis *et al.* found that the turn of the looker's head makes the observer perceive the looker's gaze as deviated into the opposite direction. Displaying the looker's face on a TV screen increased this effect. The turn of the TV screen produced a slight shift of perceived direction into the same direction into which the TV screen was turned. An observation common to all conditions was the overestimation of horizontal angles of gaze that were not directed towards the observer's eyes. This overestimation however was not observed for vertical angles. In this case, only a very slight upward deviation was measured. Figure 12 shows charts for the different conditions, putting the gaze direction as perceived by the subjects in relation to the gaze direction intended by the looker.

The experiments of L. A. Symons, Lee, Cedrone & Nishimura (2004) are equally based on the previously described experiments but with the aim to investigate the influence of different factors on the perception of gaze. In detail they investigated the effects:

- position of fixation targets relative to the subject (in the plane of the observer's face or not - dyadic vs. triadic eye gaze),
- visibility of the movement of the eye during the saccade towards the fixation target,
- visibility of both or only one eye of the looker and
- presentation of the eyes as a real physical stimulus or as an image presentation only.

When comparing static gaze to dynamic gaze Symons *et al.* found no significant difference of acuity. In the case of dynamic gaze, the looker first fixates the observer and than executes a saccade towards a gaze target. The movement of the saccade as well as the fixation are visible to the observer. In the case of static gaze, the observer only sees the looker's fixation when

the target is reached. Symons et al. use a different criterion to determine the acuity of estimation of gaze direction than Gibson & Pick. They describe it as ‘the 75% point on the psychophysical function’ (Gescheider (1985) cited in L. A. Symons *et al.* (2004)). The acuity of gaze perception they determined corresponds however to the previously reported measurements once it is converted to the visual acuity as used by Gibson & Pick. The values of acuity determined by Symons et al. are 0.5’ minutes of arc of visual angle for static and 0.52’ for dynamic gaze. The difference between the two is not significant. The acuity they measured for the perception of gaze from a single eye only was significantly lower and corresponds to 0.7’ of visual angle. The direction of gaze was overestimated towards the side of the visible eye.

For the estimation of gaze direction based on the presentation of video images of a looker, they measured a mean acuity of 0.4’, which is comparable to a live gaze stimulus. However, the resolution of the image influenced the acuity of gaze perception and acuity is decreased for angles away from direct gaze.

Svanfeldt, Wik & Nordenberg (2005) tested if the gaze of an animated head displayed on a computer screen can be perceived as eye-directed and how precise this perception is. They used static presentations of a MPEG-4 animated head under different angles of gaze and head orientation (see Figure 13). They found that almost 30% of the gaze was interpreted as eye-directed. This is more than the percentage of gaze actually intended to show fixations of the eyes. Only when adding the images showing gaze directed slightly to the right or left of the center, the number of stimuli corresponds to the measure of 30%. In the distribution of positive answers by the subjects, an offset towards one side was observed. This may be due to inaccuracies in the texturing and shaping of the eye region. It is not explained to what extent the position of the subjects relative to the screen was controlled. According to Anstis et al. this may have an influence on the perception. The offset found for the stimuli with the head slightly turned is due to an overestimation of the angles between head and gaze orientation and agrees with the findings of the above-cited studies.

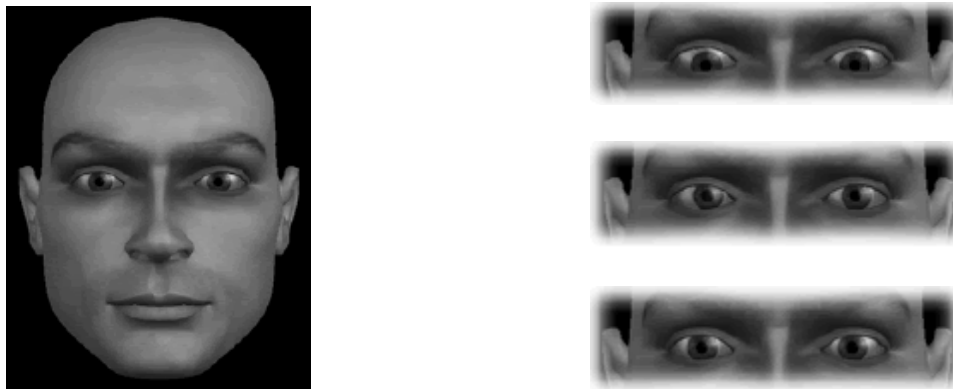


Figure 13: MPEG-4 animated head used in Svanfeldt *et al.* (2005) and three examples of eye region for accommodation to different distances.

Noll (1976) investigated whether the perception of gaze direction associated with head turn depends on which eye is looked at. He chose a scenario similar to the one described by Gibson & Pick but with black and white photos used as stimuli. He used combinations of head orientations and gaze targets straight or at either side. In addition, each of these combinations was presented with both or only one of the two eyes visible (masking the other eye). He also found an overestimation of the angle between gaze and head orientation. This overestimation is opposite to head turn as already reported in the previously studies discussed. However, Noll’s experiments confirmed that this effect depends on which eye is looked at. According to his results, the eye that is closer to the observer generates the overestimation of

gaze angles. In contrast, the observation of the other eye generates a perception that corresponds better to the actual direction of the looker's gaze.

Chen (2002) used video recordings of different lookers with the aim to delineate the range of fixation angles that produce perception of eye-directed gaze. He found that straight gaze and gaze directed to targets directly below the eyes had a high probability to be judged as eye-directed gaze. As an explanation Chen proposes the change of position of the iris with respect to the sclera that, compared to other possible directions, is minimal when changing gaze from straight ahead to downwards. This effect was independent of eye color and contact lenses or glasses. The quality of video signal makes no difference, but when a closed interaction loop between the looker and the observer is established, like in video conferencing, the subjects appeared to be more tolerant. Chen explains this with a bias produced by increased expectations of the subjects. He assumes that subjects involved in such a mediated interaction expect to be directly looked at, just as they would be in normal face-to-face interaction. He proposes that there is a critical visual angle for the perception of eye contact and that this angle is influenced by the expectations of the observer. In contrast to Anstis et al. they found no tendency of subjects to judge gaze with small downward oriented angles as straight. They rather described the perception of vertical angles as veridical. However, they did not ask the subjects to indicate when they had the impression of eye contact. Such a suggestive question may influence the subjects' answers and provoke a bias. Chen concludes from his results that in order to enable the perception of eye contact in video conferencing, the eyes should appear on the display in an angle that does not exceed 5° relative to the camera from the user's point of view.

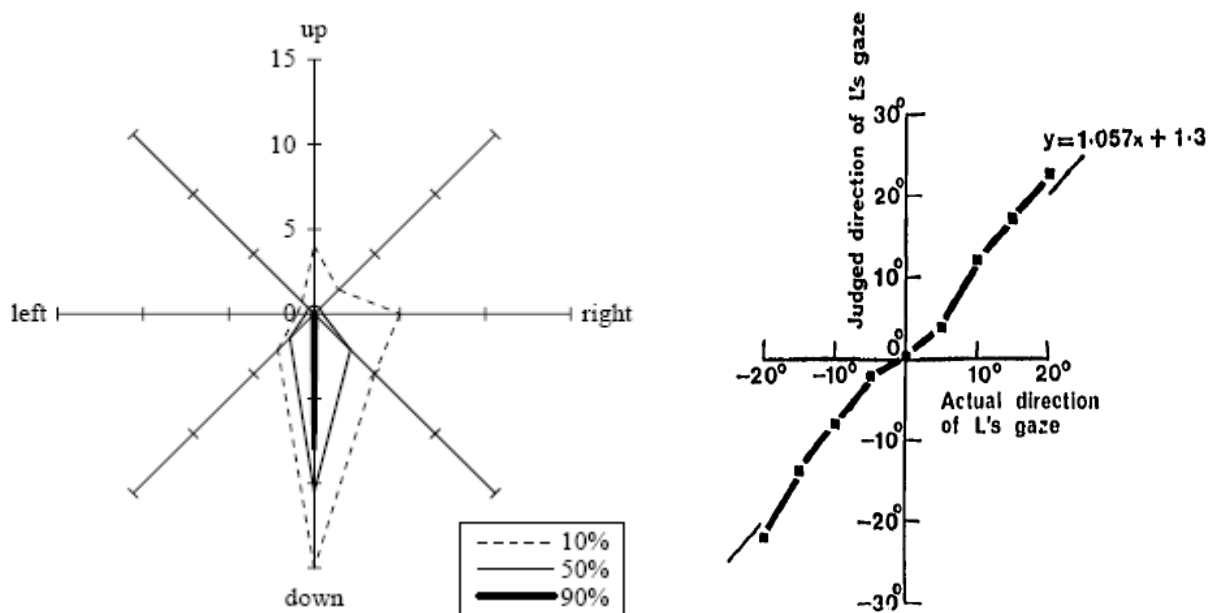


Figure 14: Left representation of angles of gaze direction of the looker relative to straight ahead. The lines delineate ranges of angles that produced the perception of eye-directed gaze at a certain percentage (measured by Chen (2002)). Right: Relation of perceived and actual angle of vertical gaze direction as measured by Anstis *et al.* (1969)

In the study by Svanfeldt *et al.*, Gibson & Pick as well as in the study by Chen the subjects may have been influenced by the task to report cases of eye-directed gaze. Such a definition of task is susceptible to condition the subjects and may result in more positive answers than in a task that does not propose a preferred response. This is for instance the case in the experiments by Cline and Anstis *et al.*. They offer multiple targets for the estimation of the

looker's gaze. The fact that the different experiments deliver similar results however confirms the respective conclusions in spite of the different instructions given to the subjects.

Gamer & Hecht (2007) propose that mutual gaze should be considered as what they call a 'cone of gaze' rather than the direction of a single ray. They developed an experimental setup where subjects could adjust the gaze direction of a virtual head or a real looker until they felt looked at or not anymore according to different start configurations. They found a constant gaze cone with an angle of 9.3° for a distance of 1m between the subject and a virtual head and 8.2° for a distance of 5m. Their results confirm the influence of head orientation on the overestimation of gaze direction. They equally confirm its influence on the perception of mutual gaze when the gaze direction actually lies between the head direction and real eye-directed gaze. They further report that at short distances (1m), the width of the gaze cone was not significantly different between a real face or a virtual face displayed on an ordinary 2D display or a 3D display.

1.3.2.2 *Important features for the estimation of gaze direction*

Several of the studies discussed in the previous section speculated about the visible properties of the eye that may be decisive features for the perception of another ones gaze direction, namely the visible parts of the sclera around the iris.

Noll (1976) found that the two eyes have a different impact on the estimation of gaze direction, depending on their laterality relative to straight ahead direction. This is of special importance in the context of overestimation of gaze angles when the head is turned relative to gaze direction. Several researchers argued that the visible parts of the sclera have a crucial influence.

Ricciardelli et al. performed an experiment investigating in detail the relation between perception of gaze direction and shapes of the visible sclera and iris. For this purpose, they changed the eye region of grey scale portraits to a black and white coloring of iris and sclera (see Figure 15 and Figure 16). The eyes were presented either with the initial positive or with inverted (negative) contrast. Presentations included all combinations of gaze directions (straight, 30° to the left and 30° to the right), head direction (straight, 30° to the left and 30°) and representation of contrast of the eyes (positive or negative). This results in 18 different stimulus images (see examples in Figure 16).

The results show a large and significant influence of contrast polarity on perception whereas negative eye stimuli produced worse results (52.3% correct overall versus 93.5% for positive contrast). Negative contrast of eyes impaired performance for every case except direct gaze in a straight face. This is a special case, as the appearance of the eye is completely symmetrical. It produced the largest impairments when the eyes and the head were both deviated in the same direction. This condition produces many erroneous 'straight' responses.

Ricciardelli *et al.* (2000) consider that the observed effects are probably due to the inconsistency between the negative contrast of the eyes and positive contrast of the remaining face. To clarify this question, they conducted a similar second experiment where the contrast polarity of the face was varied, too (see Figure 15, right).

The results of their second experiment confirm the effect of the polarity of contrast of the eye region. This effect is therefore independent of the polarity of the remaining parts of the face. The effect of polarity of contrast occurred even when other colors (dark red and light green) were chosen for the representation of sclera and iris. The mere effect of unnatural (black and white) representation of the eye region can be excluded. In contrast to the estimation of gaze direction, the estimation of head orientation is not affected by changes of contrast polarity.

Ricciardelli *et al.* (2000) conclude that the important feature for the estimation of gaze direction is the contrast of brightness between the sclera and the iris. They suggest that there

is a specialized neural system dedicated to the treatment of eye-stimuli, relying on the contrast between the darker iris and the light sclera.

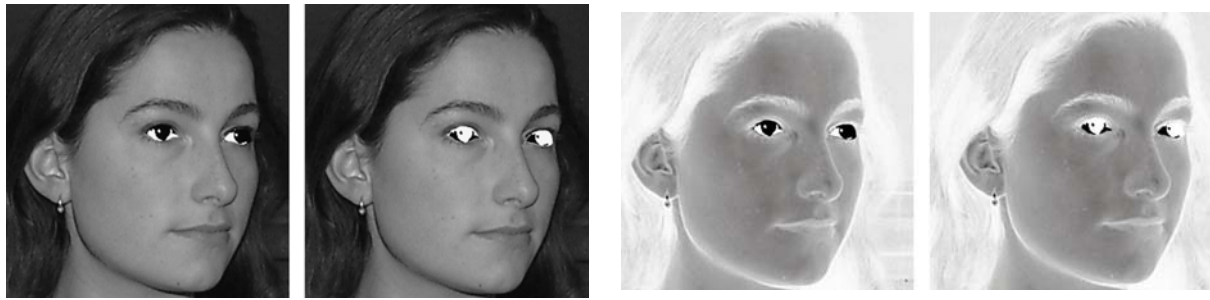


Figure 15: Examples of stimuli, with face as positive contrast representation (left) and negative contrast representation (right), both combined with positive and negative contrast for the representation of the eyes. In all four images, the head and gaze are both directed 30° to the observer's right (Ricciardelli *et al.* (2000)).

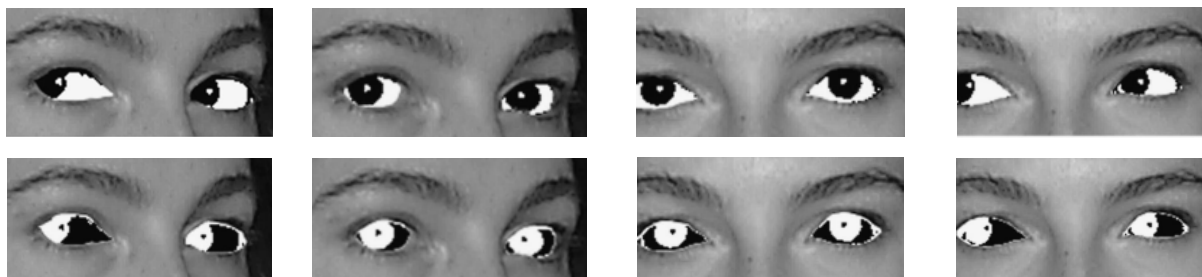


Figure 16: Further examples of stimuli as different combinations of head and eye orientation, as cutout of eye region only. From left to right: eyes-left with the head facing right; eyes-straight with the head facing right; straight eyes in a straight head; eyes-left in a straight face. Top row: positive contrast. Bottom row: negative contrast (Ricciardelli *et al.* (2000)).

1.3.2.3 Discussion

The estimation of gaze direction of another person results from a joint evaluation of head orientation and relative orientation of the eyeballs. The acuity of this estimation is of about ≤ 1 minute of arc, and varies depending on the relation between relative orientation of eyes and head as well as on absolute orientation of gaze.

When eye and head orientation are not aligned there is generally an overestimation of gaze. This results for instance in the impression of being looked into the eyes, although the looker's gaze is directed to the subject's ear while the head is orientated towards the subject's shoulder of the same side. This effect is however not observed for vertical angles. There is evidence that overestimation originates from the vergence of the eyes, due to which the angle of rotation is not identical for the two eyes. Whether there is overestimation of gaze direction depends on the eye that is fixated. In general the most important feature for the estimation of gaze direction is the contrast of brightness between the sclera and the iris and how the bright sclera surrounds the darker iris.

A general consensus of the discussed studies is the judgment of eye contact as a special event. However, the impact of eye contact and mutual gaze on cognitive processing has to our knowledge never been investigated in detail. We hypothesize that there is a special cognitive treatment related to the occurrence of mutual gaze. In subjective evaluations, subjects may classify gaze under different angles and with different conditions of presentation as mutual gaze. There is no evidence that the perception and cognitive processing of photos or videos is identical to real eye contact as it occurs during real face-to-face interaction.

Very important information in the context of our own experimental work is the fact that these characteristics of gaze perception are not considerably altered if the faces that serve as stimuli are displayed on a screen, even when synthetic faces are used.

1.3.3 Gaze direction, deixis and visual attention

We have seen that gaze is an efficient deictic cue, already perceived by young infants. Several studies have investigated in detail how shifts of attention may be initiated with special cues, particularly with human faces or eyes. These experiments also show that the direction of visual attention does not automatically coincide with the direction of gaze. The following sections discuss eye gaze as a deictic stimulus as well as the relation between attention and gaze direction.

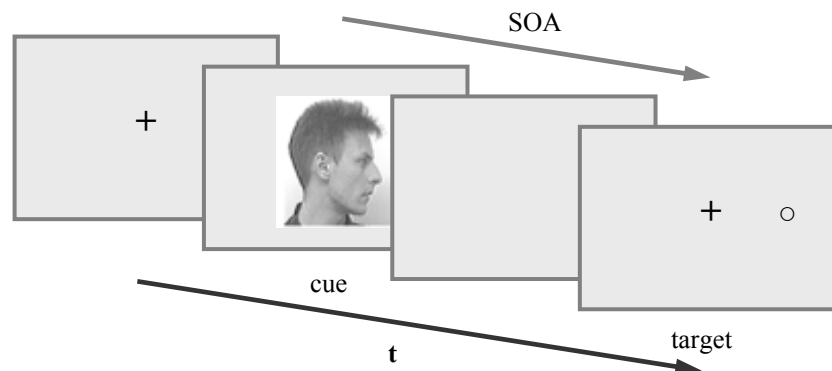


Figure 17: Schematic representation of the sequence of displays used by Langton, Watt & Bruce (2000) according to the Posner paradigm. Subjects are asked to fixate the cross in the middle of the display. An image of a face is used as cue and a circle as target. The time span between the onsets of these two stimuli is the stimulus onset asynchrony (SOA).

1.3.3.1 Deictic capacity of eye gaze

Scenarios for the analysis of deictic properties of eye gaze are commonly based on the Posner paradigm (Posner (1980) cited in Langton *et al.* (2000)). Figure 17 shows a schematic representation as an example for the sequence of displays used in such scenarios. It consists usually of four phases. At first, the attention of subjects is focused to the centre of a screen asking them to fixate a marker in its middle. Then a priming cue is shown, a few hundred milliseconds before a target symbol is displayed. The experiment varies the distributions of cued or uncued locations on the screen with different delays between the cue and the target. This delay is denominated the ‘stimulus onset asynchrony’ (SOA). The subjects are asked to indicate the appearance of the target as fast as possible and reaction time is measured. The impact of priming cues is directly connected to reaction time. The studies that use eye or face stimuli as priming cues, are of great interest for our experimental work.

Driver, Davis, Ricciardelli, Kidd, Maxwell & Baron-Cohen (1999b) used a picture of a human face as cue, that was either oriented to the right or the left. The subjects were given the task to discriminate between two symbols as targets (letters T or L), which appeared randomly at the left or right side of the screen, independently from the cued direction. For SOA of 100ms, 300ms and 700ms, shorter reaction times were observed for targets appearing at cued locations. When the target was announced to appear at the uncued location, this effect was reversed and reaction time was shorter for 700ms of SOA.

The original results by Langton *et al.* indicate that there is a ceiling effect for SOA. For a SOA of 100ms they observed significantly reduced reaction time for cued direction compared to uncued locations. For a SOA of 1000ms no such difference could be observed.

Friesen, Ristic & Kingstone (2004) tested how visual attention is influenced by gaze cues compared to arrow cues. They compared the priming effect of a schematic face with an arrow. For their experiments, they distinguished 'predicted', 'cued' and other directions. The 'cued' direction is where the face or arrow points. Subjects were told that most likely the target would appear at the opposite to the cued direction. This is hence considered as the 'predicted' direction. The remaining directions are neither 'cued' nor 'predicted'.

Friesen et al. found a fast reactive response for SOA between 100ms and 600ms. The instruction that the target was most likely to appear at the opposite to the cued direction, resulted in a reduced reaction time for SOA of 600ms, 1.2s and 1.8s. There seems to be an increased reactive attention in cued direction within an interval shortly after stimulus onset and a volitional increased attention at a predicted location within an interval beginning later after stimulus onset. These two intervals of SOA overlap and do not appear exclusively. Using an arrow instead of a schematic face did not result in reactive shifts of attention when short SOAs were used. The arrow was able to direct attention only when high SOAs were used. This resembles to the effect observed with large SOA in the experiments in which predicted and cued direction did not coincide.

In the discussed experiments, authors did not specify whether the shifts of attention, documented by reduced reaction times, were accompanied by gaze shifts. In addition to reaction time Friesen et al. monitored the gaze direction for some of their subjects. These experiments showed no co-occurring gaze shifts. Shifts of attention that are not accompanied by gaze shifts are denominated 'covert shifts of attention'. Such cases where the direction of attention and the direction of gaze are not congruent are discussed in more detail in the following section.

From this section it can be retained that attention is shifted into the direction where eyes or faces point during a short interval after their presentation (up to about 600ms). This is an involuntary, reactive shift of attention. After this interval, the effect of reactive shift of attention decreases and attention can be directed voluntarily.

1.3.3.2 Gaze direction and visual attention

Modern eye tracking techniques offer various equipments to determine gaze direction for very diverse requirements. There is however no possibility to deduce the direction of visual attention from gaze. A common assumption is that the focus of attention coincides with the area where gaze is directed. Both are surely closely associated but nevertheless are not implicitly the same. There are for instance shifts of attention that are not accompanied by gaze shifts. These are denominated '*covert shifts of attention*'. They facilitate the processing of selected items without shifting gaze direction towards them (Godijn & Theeuwes (2003)). If a gaze shift is included, it is considered as an '*overt shift of attention*'. The shifts of attention resulting from the presentation of eye or face stimuli as discussed in the previous section are for instance covert shifts of attention.

A study by Paré, Richler, ten Hove & Munhall (2003) provides better understanding of visual attention during audiovisual speech perception. They investigate the influence of gaze direction on the perception of McGurk effect (McGurk & MacDonald (1976)). The McGurk effect is observed when simultaneous visual and acoustic presentations of speech do not convey the same information. This occasionally results in the perception of other sounds than either of the two that are presented with different modalities. A typical example is the presentation of a visual 'aga' along with an acoustic 'aba' which may be perceived as 'ada' or 'ava'. Paré et al. found no significant interrelation between the McGurk effect and the direction of gaze towards the mouth or the eyes of the video stimulus. The restriction of subjects' gaze towards different targets on the video presentation, such as the mouth or the eyes, did not result in a significantly different rate of the McGurk effect. Only fixations at

about 10° or more above the mouth resulted in significantly less occurrences. This means that there is a rather large area of the face around the mouth that can be fixated without interfering with the perception of the McGurk effect. Conclusions on the visual speech perception in general can however not easily be made. The McGurk effect may result from the interpretation of relatively coarse movements of the jaw, compared to lip and tongue movements. The increase of cognitive load associated with the distance between gaze target and the mouth is not clarified either. The results show that vision is active during audiovisual speech perception. Given a natural scan path, comprising the eye and mouth area of the face, it can be assumed that qualitative changes in the eye and mouth region can be detected and that even visual speech perception may be possible, independently of actual gaze direction.

This study shows that increased attention may be directed at locations that do not coincide with the foveated target and that information may still be retrieved. On the other hand, there are studies that show that attention is crucial to consciously perceive any information, even when present in foveated locations. Events occurring at foveated locations are ignored, unless we are actually attentive to them. Items that appear in a scene may be ignored when not corresponding to the current task, even when they are salient or lasting.

Simons & Chabris (1999) recapitulate several studies reporting these effects that they confirm with their own experimental work. For instance, they produced a video showing two teams wearing different colors and passing basketballs between the members of a team. Subjects were asked to count the passes between members of the same team (white or black shirt). Right in the middle of the video, a woman with an open umbrella or a person in a gorilla costume passed through the group of people. An important part of the observers did not remark the masqueraded person.

The results reported by Simons & Chabris prove that objects that are not expected may be ignored, even when they appear at fixated areas. This effect is denominated '*inattentional blindness*'. A similar effect describes events, where changes to objects are ignored that occur during a temporary interruption of view or during a saccade. These are examples of '*change blindness*'.

We conclude that gaze direction does not automatically coincide with direction of attention. Attention may be directed to an object in the peripheral visual field that is actively perceived, although another object is fixated. Furthermore, attention is required to perceive actually any object in the visual field.

1.3.4 Gaze patterns during examination of human faces

In the previous sections, we discussed gaze as a particular exogenous stimulus and its impact on the perception of observers. In the first place, however, the orientation of eyes is controlled for scrutinizing the visual scene. The most interesting scene stimuli in the context of our work are human faces. In the following, we discuss the gaze patterns observed during the examination of human faces and the factors that may influence these patterns.

Eyes and faces are of special importance to humans, which can be shown already with very young infants (see section 1.3.1, page 28). Vatikiotis-Bateson, Eigsti, Yano & Munhall (1998) investigated the gaze behavior of subjects looking at faces. They conducted experiments to analyze gaze patterns during audio-visual perception of monologues. In a pilot study comparing audio-visual and audio-only presentation of long conversational monologues, they found that intelligibility of noise-masked speech could be increased by the presentation of the corresponding facial movements. In their main experiment they investigated the influence of noise and image size on gaze patterns.

To classify and analyze the data they defined three zones on the displayed face, two for the eyes, separated by the midline of the nose, and one for the mouth, which covered the lower

half of the face including the tip of the nose. They measured duration of fixation and saccade frequency.

For increasing noise, Vatikiotis-Bateson et al. found that duration of fixation in general increased and that the proportion of fixation time dedicated to the mouth, increased from about 35% to 55%. At the same time, they observed a decrease in frequency of saccades between eyes and mouth of up to 50%. Variations of image size from life size to five times life size show no major influence on the gaze behavior. However, saccades between the mouth and eyes become more frequent and a tendency to prefer one eye develop. Saccades between eyes and mouth tend to cross the midline of the face. No influence of phonetic identity on gaze direction could be found.

Lansing & McConkie (1999) tried to find relations between task and gaze patterns during the examination of speaking faces. They chose the visual presentation of two-word sentences. Before the stimulus presentation, subjects were given different instructions. They are asked to distinguish between two proposed alternatives concerning the word content of the sentence, the intonation (question or statement) or the primary sentence stress on first or second word. Lansing & McConkie found that the gaze direction of the subjects varied depending on task and time course of the sentence. When subjects were attentive to segmental information (word content), they showed a stronger tendency to gaze towards the mouth compared to cases where they had to be mainly attentive to prosodic information which resulted in more frequent gazing at upper parts of the face such as the eyes and eyebrows. In a second experiment, they modified the visual stimuli. They masked the face with a still image that only unveiled the parts of the face below the nose. This perturbs the detection of information concerning intonation compared to a full-face presentation. In contrast, detection of stress and segmental information are not affected.

From these results, Lansing & McConkie conclude that information from the mid- and lower face region is sufficient for the distinction of word and sentence stress, whereas information on intonation is distributed more broadly. In general, they take their results as evidence that gaze direction can indicate where subjects search for facial features. The observed gaze patterns confirm qualitatively the results by Vatikiotis-Bateson, Munhall, Hirayama, Kasahara & Yehia (1996) concerning the eyes and mouth as dominant gaze targets along with central parts such as the nose. The exact proportions measured however were not identical. This is not astonishing as the study by Vatikiotis-Bateson et al. used longer monologues. Furthermore, Lansing & McConkie instructed the speaker explicitly to use facial expression to differentiate intonation and stress patterns. This has probably led to exaggerated facial movements that would not appear during ordinary conversation.

1.3.5 Gaze direction in text reading

For completeness, we will shortly mention the characteristics of visual perception during reading, although this is of no major importance for the work reported here.

During reading gaze moves back and forth between words. There is about 85% of forward movement, from left to right, and 15% of backward movement. Saccades are typically directed to an area between the beginning and the centre of a word and of an amplitude of about seven to nine characters (Liversedge & Findlay (2000), for English readers). Saccades usually skip short function words. Fixations during reading have a duration of 60 to 500ms with a mean of 250ms. During fixations letters are both perceived to the right and to the left of the fixated point, whereas perception is extended in reading direction. Saccades and fixations of course are influenced by text processing and word recognition.

1.4 EYE GAZE AND SOCIAL INTERACTION

Gaze and eyes have been found to have an outstanding meaning in human perception. As a matter of course, they also play a major role in human interaction and communication. Butterworth (2003) discussed the importance of pointing in combination with speech for developing the capability of linguistic referencing of objects and how pointing, speech and gaze are related in the establishment of joint visual attention and grounded communication. These aspects of gaze seem to be quite general across culture and language barriers. It is obvious that among the different modalities of communication eye gaze plays a particular and exceptional role. It may convey meaning and information by itself but also influence and alter information that is transmitted by speech or other modalities of communication. In the following, we will discuss gaze in the context of social interaction. Several researchers made an effort to describe and quantify the communicative value of gaze and the large variety of communicative functions that gaze may adopt.

Kendon (1967) names several distinct communicative functions of gaze in face-to-face interaction, such as:

- regulating conversational flow
- providing feedback
- communicating emotional information
- communicating the nature of interpersonal relationships
- avoiding of distraction by restricting visual input

In the work by Argyle & Dean (1965) and Argyle & Cook (1976) the functions of gaze in the regulation of conversation are further:

- feedback at the end of speech about how this has been received
- looking away at the beginning of speech or while thinking for an answer of a question, for example in order to avoid extra input from eye contact that would be distracting (several studies indicate that verbal questions produce gaze shifts to the right and downwards, and that spatial questions produce shifts to the left and upwards)
- the mere signaling that the channel is open and interaction may proceed
- eye gaze as a sign to give the word to another
- eye contact as obligation to interact

They also assign an important influence to gaze in the context of social relation and status, such as:

- friendship
- hate
- dominance (0.69 correlation between amount of gaze and evaluation of leadership (Burroughs, Schultz & Autrey (1973) cited in Argyle & Cook (1976))
- submission
- attractiveness (pupil dilation that may influence the attractiveness of a person on photographs (Hess (1965) cited in Argyle & Cook (1976))
- evaluation of personality (increased gaze may lead to characterization of a person with more positive attributes (Kleck & W. Nuessle (1968), Cook & Smith (1975) cited in Argyle & Cook (1976)) gaze aversion led to a characterization with negative attributes)
- sexual attraction etc.

Kleinke (1986) also cites several authors confirming these observations.

1.4.1 Definition of terms concerning gaze

There are different terms that may be used to express different aspects of looking. Kleinke (1986) proposes definitions for the different terms used to describe gaze in the context of interpersonal interaction. He defines 'gaze', 'looking', 'glance' and 'staring' as describing gaze that is directed to another person's face, whereas 'staring' is gaze that is maintained independent of the other person's behavior, as if it was an object. With 'mutual gaze' and 'eye contact', he names the simultaneous looking at each other of two persons. 'Gaze avoidance' is the intentional avoidance of eye contact, whereas 'gaze omission' is merely the failure to look, but without the intention of avoidance. These definitions correspond mainly to the common use of these terms in the literature. There is however, a distinction that may be made between mutual gaze and eye contact. Mutual gaze may include gaze that meets the face but not directly the eyes. Eye contact in contrast necessitates both persons to direct gaze at each other's eyes. This distinction is often not clarified in literature, which may be due to the difficulty or impossibility to distinguish between these two events with the formerly available technical means. Only with recent eye tracking techniques and devices such a detailed distinction is possible during the measurement of gaze direction. Nevertheless, even if the technical equipment allows for this distinction, it is not clear, whether mutual gaze and eye contact are different in the sense of a special cognitive reaction. There may be neural reactions beyond the mere recognition of gaze direction in this case. The studies discussed in section 1.3 (page 28) however suggest the existence of specialized neural processes. Baron-Cohen (1995) discussed such dedicated neural processes in the context of detection of gaze direction of another person in general (see section 1.4.2, page 45).

Kampe, Frith & Frith (2003) conducted an experiment in order to localize areas of the brain that are involved in the processes of building hypotheses about another persons mental states which they describe as 'mentalizing' (see also next section). They measured significant brain activity in certain regions when subjects are confronted with the self-referential signals such as hearing their own name or with prolonged eye contact. Such prolonged eye contact was generated by color images of faces as stimuli that looked straight ahead. A similar experiment monitoring the gaze direction of both subjects might be interesting to investigate under which conditions the special sensation of eye contact is experienced. It may clarify whether during face-to-face interaction this sensation is restricted to gaze that is directed to a limited area around the eyes, or if it extends to a larger region of the face exploiting peripheral vision. The cone of gaze as proposed by Gamer & Hecht (2007) partly responds to this question (see section 1.3.2.1, page 32).

1.4.2 Eye gaze and theory of mind

A valuable approach to understand the special functions of gaze in social interaction and their origin is provided by the work of Baron-Cohen (1995). He summarizes his own work backed up by studies of numerous other authors in psychological research on brain activity, visual perception and social interaction. He puts forward a well-established theory on important mechanisms of human social interaction and their underlying mental processes. In a similar way as proposed by Butterworth (2003), he describes eye gaze as a crucial component for the development of audio-visual communication (see section 1.3.3, page 40). He proposes a comprehensive 'theory of mind' that he sets up as a modular system.

Baron-Cohen proposes two modules - an 'intentionality detector' and an 'eye direction detector' - as the basis of his 'theory of mind'. Both are available very early in infancy.

Reactions associated with these two modules can be observed during the first months after birth.

The ‘intentionality detector’ uses any input modality such as vision, audition and touch to attribute primitive volitional states like goals and desires to anything that may be considered as an agent. It assumes any perceptible event as caused by an agent and tries to attribute intention to the actions of this agent. Observations made with certain brain lesions and experiments on brain activity with monkeys are cited as evidence that this is a localized module that can be dissociated from other parts of the cognitive system.

The ‘eye direction detector’ only uses vision and is considered as specialized part of the human visual system. It has the three basic functions:

- detection of eyes or eye-like stimuli
- determination whether the gaze of another person is directed towards the observer or somewhere else
- generation of the knowledge that another person perceives the object he or she looks at

Based on the observation of behavioral changes after brain lesions of human patients and on experiments on brain activity of monkeys the ‘eye direction detector’ (EDD) module is assumed to be situated in the superior temporal sulcus and in the amygdala.

Baron-Cohen describes the inferences produced by these two modules as of a dyadic nature. They can only specify the relations between two entities such as an agent and an object. They do not allow for grounded interaction between two individuals, where both are paying attention to the same item. This is based on triadic representations that specify the relation between the self another agent and an item.

The ‘shared-attention mechanism’ (SAM) is a third module dedicated to triadic representations. It combines the dyadic representations of the own perceptual states with the assumptions about another person’s dyadic representations. Furthermore, it also integrates the volitional states that the ‘intentionality detector’ (ID) assigns to observed items. The triadic representation needs to be affirmed through repeated confirmation. Such actions of confirmation and verification using vision can be observed in children from the age of about 9 month (see also section 1.3.1, page 28). This is the age at which pointing gestures emerge. Although the ‘shared-attention mechanism’ module may exploit any modality, visual perception is the most important. It provides probably the fastest and easiest method of confirmation and functions also over distances that impede touch or speech.

The top-level module as described by Baron-Cohen is the ‘theory of mind mechanism’ (ToMM). It adds the distinction between assumptions of another person and the own assumptions to the previously discussed attribution of volitional and perceptual states. These assumptions may be different about the same object. The ‘theory of mind mechanism’ incorporates mental states such as pretending, thinking, knowing, believing, imagining, dreaming, guessing and deceiving.

The mind reading system as proposed by Baron-Cohen, is an appropriate illustration to represent the importance of visual perception for the social interaction of healthy humans and to enable a comprehensive understanding of the emerging functions of gaze. He describes the abnormal social behavior observed with autistic persons as ‘mindblindness’, which he ascribes to the dysfunction of the shared attention mechanism and the theory of mind mechanism. The consequences of these dysfunctions evidence the importance of eye gaze and visual perception for successful social interaction.

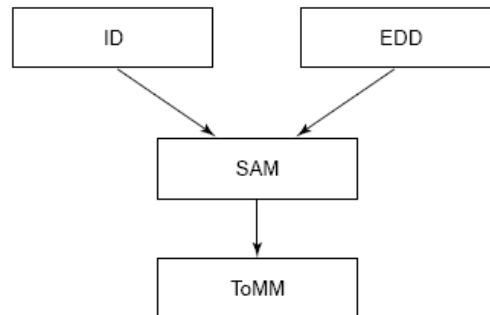


Figure 18: Schematic diagram of the relationship between the four components proposed by Baron-Cohen's for the modeling of the human mind-reading system.

1.4.3 Functions of gaze in conversation and dialogue

1.4.3.1 Adam Kendon's pioneer work

In the context of discourse and dialogue, Adam Kendon's pioneer work on the direction of gaze in social interaction has an outstanding status of publicity (Kendon (1967)). It serves as an important reference and basis for several approaches to model gaze direction for virtual characters (see chapter 'Modeling of gaze and social interaction', page 64). His work is not the first research attending to gaze, but one of the first to analyze in detail the function of gaze in social interaction. Argyle & Cook (1976) report that there was almost no research on gaze and social behavior until the early 1960s. The work by Kendon documents direction of gaze, facial expression, position of head, hands, arms and trunk of subjects getting to know each other in an unrestricted but monitored face-to-face conversation. The events were registered as pictographic symbols in a list along with a transcription of the speech related to each frame (see Figure 19). A mirror was used for capturing of both subjects in one photograph every 500ms for later analysis. A synchronization signal from the camera was recorded along with the sound.

Kendon considered that gaze has two main functions for the looker. He assumed that on the one hand it is an act of perception, by which one interactant can monitor the behavior of another. On the other hand, he supposed it to be an expressive sign and regulatory signal by which one may influence the behavior of another.

The most important and meaningful result is that there are different amounts of gaze depending on whether a subject is speaking or listening. In Kendon (1967), these states are however not clearly defined and listening intervals seem to serve more as a distinction between speaking and not speaking. They probably include also mutual silence, thinking or other activities. For 80% of the subjects, less than 50% of gaze was reported to be directed to the interlocutor while speaking. The amount of gaze varied between 20% and 65% depending on the subjects. While listening, 64% of the subjects directed more than 50% of the gaze to the interlocutor, varying between 30% and over 80% for the different subjects. While listening, gazes towards the interlocutor tend to be relatively long, interrupted by very brief aversions of gazes. In contrast, during speaking, averted gaze lasts longer and accounts the same duration as gazes towards the interlocutor. Kendon reports that mutual gaze occurs only during short periods of about 1s. A more detailed analysis shows that the duration of these intervals and their periodicity is speaker dependent. One of the subjects for instance tends to avert gaze sooner and more often.

Kendon found that for each subject of a dyad, gaze durations and frequency of changes of direction were correlated. In particular one subject that participated in two interactions showed different values for these parameters adapting to the respective interlocutor.

A closer look at the gaze pattern in relation to speaking showed a common strategy that for most individuals consisted in looking away at the beginning or before a long utterance and looking at the interlocutor when approaching the end. Kendon proposed the need for concentration and the attempt to inhibit a take over by the other as probable reasons for such aversion of gaze. Furthermore, he considers the checking of attention and comprehension of the listener, the will to take the turn or to wait for comment or the intent to leave the turn to the other as possible motivations to look at the interlocutor. If for instance the looking up at the end does not occur, delays or omissions of turn taking by the other are observed in several cases.

Fluency of speech and gaze seem to interact. During hesitation, fewer gazes towards the listener are observed. When the speech is fluent, gaze increases. Similarly, the speech rate tends to be higher when accompanied by gaze towards the interlocutor. Utterances with positive attitude or affect and utterances signaling attention, such as humming noises, are also accompanied by gaze towards the interlocutor. Negative utterances or short utterances of agreements tend to coincide with short gazes away from the interlocutor.

This work by Adam Kendon is well documented, uses a relevant amount of data and provides a very detailed transcription of events. It must however be noted that due to the technical means available at the time, the sampling rate is rather coarse. At a frame rate of 500 ms for sampling of visual appearance, no reliable observation of saccades is possible since in such an interval multiple saccades may occur. Furthermore, the data is analyzed by a human observer based on photographs, which does not permit the establishment of a reliable relation between gaze direction and possible targets. Therefore, no fine assignment of fixations to targets or regions of interest is possible that would allow for a finer grained distinction than between fixations directed to the face region or elsewhere. This has also to be kept in mind when interpreting the assignment of mutual gaze.

	NL SPEECH	EYES	BROWS	MOUTH	HEAD	GAZE	GAZE	HEAD	MOUTH	BROWS	EYES	JH SPEECH
352	and um	○	∩	○	□			□	-	W	∩	
3	sometimes	○	∩	○	□		■	□	-	W	∩	
4	of course it's	∩	∩	○	□			□	∩	∩	∩	
355	only one of	∩	∩	○	□			□	=	∩	○	
6	parents in which	∩	∩	○	□			□	=	W	∩	
7	case you can	∩	∩	○	□		■	□	=	W	∩	
8	take it	∩	∩	○	□			□	=	W	∩	
9	away and	○	∩	○	□		■	□	=	W	∩	
360	let the	○	∩	○	□		■	□	=	W	∩	
1	other one feed them	∩	∩	○	□			□	=	W	∩	
2		○	∩	○	□			□	=	W	∩	
3		○	∩	○	□			□	=	W	∩	
4	itself	○	∩	○	□		■	□	=	W	∩	
5		○	∩	-	□			□	⊙	W	∩	
6		○	∩	-	□			□	⊙	W	∩	some breed-
7		○	∩	-	□			□	⊙	∩	∩	ers
8		○	∩	-	□			□	⊙	∩	∩	un
9		○	∩	-	□			□	⊙	∩	∩	pair
370		○	∩	-	□			□	⊙	∩	∩	with
1		○	∩	-	□			□	⊙	∩	∩	infer-
2		○	∩	-	□			□	⊙	∩	∩	ior
3		○	∩	-	□			□	⊙	∩	∩	birds for
4		○	∩	-	□		■	□	⊙	∩	∩	this purp-
5		○	∩	-	□			□	⊙	∩	∩	ose
6		○	∩	-	□			□	⊙	∩	∩	
7		○	∩	-	□			□	⊙	∩	∩	em I mean
8		○	∩	-	□			□	⊙	∩	∩	
9		○	∩	-	□			□	⊙	∩	∩	
380		○	∩	-	□			□	⊙	∩	∩	those that don't

Head	□	Head erect, face pointing forward.	Mouth	-	Closed, lips relaxed
	◁	Head turned left.	⊕	⊕	Lips relaxed, mouth open.
	▷	Head turned right.	⊙	⊙	Lips pouting
	↖	Head tilted left.	∩	∩	Lips drawn tight at corners.
	↗	Head tilted right.	⊗	⊗	Lips pressed forward, "pursed"
	⊥	Head tilted back.	Eyes	○	Fully open
	⌞	Head tilted forward.	⌘	⌘	Narrowed eyes.
Brows:	∩	Normal.	⌘	⌘	Closed eyes
	∩	Raised brows.	Gaze	■	p looking at q
	∩	Puckered or "frowning" brows.			

Figure 19: Example of a transcription protocol from an interaction analyzed in Kendon (1967). The visual appearance of the eyes, brows and mouth, the head position, gaze direction and the speech are annotated in intervals of 500ms. The meaning of the used symbols is given below the table.

Apart from the work by Kendon, there are several publications with similar importance in the context of gaze behavior in social interaction. Argyle & Cook (1976) summarizes functions of gaze as reported by other researchers:

- long glances used by speakers to signal grammatical breaks such as end of utterance
- speakers look up at grammatical breaks for feedback
- glances may be used to emphasize parts of an utterance or to point
- gaze of listeners signal attention, readiness to continue to listen, or encourage the speaker
- aversion of gaze signals lack of interest and disapproval
- the listener gazes at the speaker to decode facial expression and the direction of gaze

Thórisson (2002) used these findings amongst others for the animation of a conversational agent (see 1.6).

1.4.3.2 Turn taking

Novick, Hansen & Ward (1996) conducted an experiment with four pairs of subjects in a joint memory task. Subjects were seated facing each other over the edge of a table to jointly complete a sequence of characters from memory. Hand gestures, gaze direction and speech were transcribed from video recordings of the interactions. For the detection of turn changes every coherent verbal utterance, even if very short, was considered as a turn.

Novick et al. found two main gaze patterns accompanying turn changes:

- 1) One pattern consists of gazing towards the listener at the end of a turn. The listener responds equally with gaze, resulting in mutual gaze for a short time until the listener takes the turn and starts to speak. This corresponds to the gaze behavior that Kendon (1967) reported to appear commonly along with turn changes. Novick et al. associate this pattern with a smoothly proceeding conversation.
- 2) The second most used pattern during turn changes proceeds similarly, but the turn recipient continues to look at the interlocutor while starting to speak, only braking gaze later during the utterance.

In addition to inspecting turn taking events, Justine Cassell, Torres & Prevost (1999) conducted experiments examining the relation between information structure and gaze direction. They classify the discourse of each speaker into three main units. Any part of an utterance delimited by pauses is considered as a turn of the speaker. It is further subdivided in 'thematic' and 'rhematic' part. The 'theme' is defined as the part of an utterance that links it to the past discourse and specifies what the utterance is about. The part presenting a new or specific contribution to the current context is defined as the 'rheme'.

A very interesting result reported by Cassell et al., is the mandatory look away when the start of a turn coincides with the start of a theme. At the end of a turn that coincides with the start

of a rheme a similar mandatory look of the speaker towards the listener is triggered. Otherwise, the gaze behavior is still correlated with information structure with a rather high probability of around 70%. Gaze changes away from the listener coincide in 44% of cases with the start of a turn. Only 16% of changes of gaze towards the listener coincided with the end of a turn. Concerning the gaze target independently of the onset of gaze, the observations correspond to the literature that reports a strong tendency to look at the hearer at the end of a turn and not to look at the listener at the start of a turn. Around 40% of gaze changes in either direction did not coincide with any of these events.

Cassell et al. conclude that the association of turn events with information structure is a strong predictor of gaze behavior and that information structure should be considered in the investigation of gaze behavior. They point out a possible relation between the willingness to give up the turn and having conveyed the rhematic part, which contributes new information to the conversation.

1.4.3.3 Gaze and interpersonal distance

Argyle & Dean (1965) propose an equilibrium theory, which assumes that two persons involved in interaction, will negotiate a certain level of intimacy, which they will attempt to maintain. They consider that several factors contribute to the perception of intimacy and that these are adjusted by the interacting persons to generate a level of intimacy so that all of them feel comfortable with. As the most important factors they identify eye-contact, physical proximity (see proxemics in Hall (1963)), intimacy of topic and amount of smiling. Observations reported in previous publications by different researchers indicate the existence of such relations (Coutts & Schneider (1975) Ross, Layton, Erickson & Schopler (1973) cited in Argyle & Cook (1976)). The attempt to install such equilibrium in interpersonal distance between people of different cultural background may cause continuous difficulties such as for instance when a forward movement by one person is followed by a backward movement by the other. Another example of every day experience is the aversion of gaze in lifts or crowded places.

Argyle & Dean conducted a preliminary test to find a reasonable distance at which to place two persons as a basis for further experiments. They found that the distances at which a subject approaches a person having the eyes closed was significantly shorter than the distance to a person with eyes open. For the main experiment, they placed subjects at distances of about 60cm, 1.8m and 3m. To control the desired distance and to ensure that it will be maintained, they placed chairs accordingly at an angle of 90° over the corner of a table. One of the participants was a confederate that was continuously looking at the other person's face to generate mutual gaze every time the subject would look up at the confederate. A hidden human observer who notes the total amount of mutual gaze as well as the average length of glances using cumulative stopwatches monitors the subjects.

Argyle & Dean report a high consistency when comparing the independent measures of two observers that seemed to have no difficulty to detect eye-directed gaze. Only a little percentage of time however is dedicated to other parts of the head than the eyes. They found significant differences of amount of eye-directed gaze. Independent of the combination of sex of the two subjects, eye-directed gaze increases along with distance between the interlocutors (see Figure 20). The length of glances also increases with distance. For interlocutors of different sex the mean amount of gaze was significantly less compared to pairs of the same sex. For the closest positioning the subjects were more agitated, showed different behavior of distraction and avoidance of gaze along with leaning back as an attempt to increase the distance.

The authors conclude that if a certain equilibrium is disturbed by one factor this could partly be compensated by another factor. This is for instance the case with interpersonal distance and amount of mutual gaze.

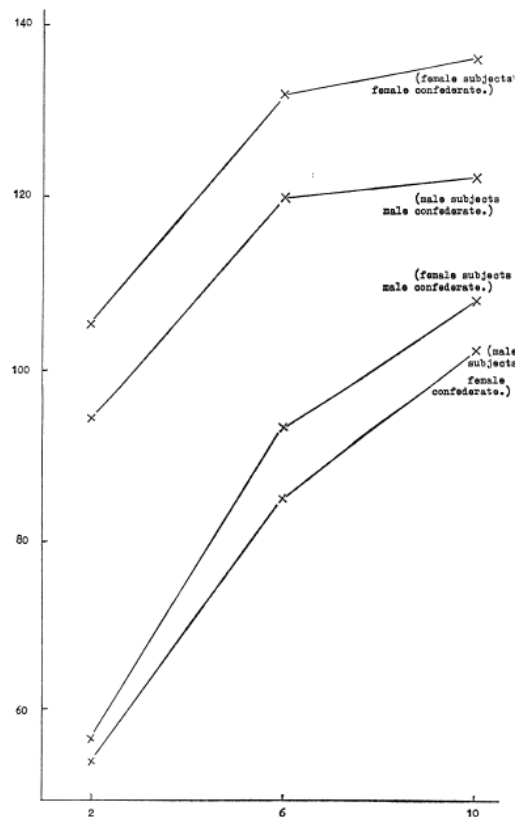


Figure 20: Total amount of eye contact in ms as measured during interactions of 3 minutes at distances of 2, 6 and 10 feet (60cm, 1.8m and 3m), and different combinations of sex of subjects. The ‘confederate’ was instructed to continuously look at the subject’s eyes in order to produce mutual gaze whenever the subject looks up (Argyle & Dean (1965)).

The validity of the equilibrium theory as proposed by Argyle & Dean has been tested in immersive virtual environments by Bailenson, Blascovich, Beall & Loomis (2001). They use immersive virtual environments that permit better control of the stimuli and the possibility to change aspects of behavior and appearance independently. They give subjects a memory task. Subjects have to approach a virtual agent in order to read and memorize information that is displayed on its front and back side. The virtual agent is animated using different levels of gaze animation and rendering of the face as well as different genders. It is placed at a fixed location in the virtual room without motion of the body apart from head movements. At the highest level of gaze animation, the agent follows the subject with head orientation and continuously staring eyes. The pupils are dilated when a subject approaches too close (see Figure 21). Apart from blinking, no further manipulation of gaze direction or other animation is conducted.

When compared to a cylindrical object the agent is treated differently by the subjects, leaving more personal space. Women, compared to men, are more likely to realize the differences in gaze behavior by the agent and adjusted the personal space they leave according to the gaze behavior predicted from the equilibrium theory. The authors explain this with the tendency of men not to maintain eye contact. They are thus less sensitive to the different degrees of animation. The performance of subjects in the memory task is degraded by the presence of the agent compared to the cylindrical object. The agent seems to absorb the subjects’ attention.

Modification of face texture of the agent as a three-dimensional model or as photograph-textured do not have any influence on the subjects.

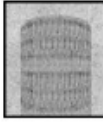








CONTROL	CYLINDER		
COND 1	EYES CLOSED		
COND 2	EYES OPEN		
COND 3	BLINKING		
COND 4	BLINKS And HEAD TURNS		
COND 5	BLINKS, HEAD TURNS, And PUPIL DILATION		

Figure 21: Schematic representation of conditions as used in the experiments in Bailenson *et al.* (2001) to verify the equilibrium theory in immersive virtual environments. The middle column informs about the characteristics of the presented stimulus and its animation. The column to the right provides the corresponding visual realizations.

1.4.3.4 Social norms and gaze behavior

Some researchers found evidence for an influence of social norms on gaze behavior. Gullberg & Holmqvist (2001) conducted experiments to analyze the effect of medium of presentation on fixation behavior. Subjects observed the narration of a story presented by person after having read it from a cartoon. The narrator was filmed at the same time to obtain a video stimulus for later presentation to other subjects on a video screen. The identical stimulus can then be used as live and as video presentation. The subject's gaze behavior is monitored with a head mounted eye tracker. Fixations are detected using temporal and spatial criterions that resemble to the dispersion-based algorithm proposed by Salvucci & Goldberg (2000) (see section 3.2.4, page 99). They show a significant reduction of face fixations in the video condition with an accompanying increase of fixations directed at immobile body parts. The authors point out a strong influence of social norms on gaze behavior. The observed differences may either be due to the missing feedback loop in the video condition that does not require confirmation by appropriate fixations or to the fact that fixations to socially less acceptable regions are not made when these may offend the looked at person. The reduction in presentation size that may have advantaged exploitation of peripheral view and therefore permitted a different scan path, could also have had an influence.

Minato, Shimada, Itakura, Lee & Ishiguro (2005), who investigated the gaze patterns of the human partner during human-human interaction and human-android interaction also conclude that gaze behavior is strongly influenced by social norms (see section 1.4.5, page 54).

1.4.3.5 *Gaze in multi-party interaction*

Vertegaal, Slagter, van der Veer & Nijholt (2001) evaluate group conversations of four persons to find out if gaze can be used as a predictor to decide whom a person is speaking or listening to. They track the gaze of one subject of each group and monitor the gaze direction, the speech activity and whom the subject addresses. While reviewing the recording of the interaction after the experiment, the subjects self-annotate to whom they intended to speak. Circular zones defined around the heads are used to automatically assign the observed fixations to the different interlocutors. They then computed the percentage of fixation time that is dedicated to the different persons during the different conversational modes.

Vertegaal et al. found that a speaker is fixated more than a listener. However, both were more gazed at than persons that had not been addressed. Speaking to all interlocutors at once resulted in more total fixations at persons than when only speaking to only one of the interlocutors. Vertegaal et al. conclude that the amount of gaze can serve as a strong indication to determine to whom a subject is paying attention. These observations confirm the findings of Argyle & Dean (1965), Argyle & Cook (1976) and Kendon (1967). Furthermore they also report that gaze directed to the faces was dedicated to the two eyes or the mouth as also reported by Vatikiotis-Bateson *et al.* (1998) and Paré *et al.* (2003).

1.4.3.6 *Summarizing the function of gaze in conversation and dialogue*

The studies of gaze behavior in conversation and dialogue confirm gaze as an important modality of communication with various functions in conversation. Argyle & Cook (1976) hypothesized that the importance of gaze in the context of regulating speech flow may be due to the fact that turn taking in dialogue cannot effectively be regulated by speech itself, as this would provoke even more turn conflicts. Visual feedback as another modality is an effective alternative as it can accompany speech.

Several studies agree about the relation between turn taking and gaze as well as about the differences in gaze behavior depending on the activity of speaking versus listening. The different steps of turn taking or course of dialogue in general, as incited coarsely as 'end of speech', 'beginning of speech' or 'pause between turns' should be clearly separated, distinguished and examined in detail. The structure of the delivered information, subdivided into 'theme' and 'rheme' is another interesting approach that is worth further examination. Furthermore, the influence of conditions of interaction used in the different studies as well as the influence of social norms is quite important. The detailed understanding of eye contact and its special neural processing is another crucial aspect of great interest.

1.4.4 Mediated interaction between humans

With mediated interaction, we mean interaction that is established via technical equipment connecting physically remote persons that are not in direct audio-visual contact. Such settings are interesting for research on virtual characters, as they restrain the appearance of human subjects to the two-dimensional working space to which virtual characters are restricted. It can unveil problems that arise from the limitation to a two dimensional appearance or unexpected effects that may interfere with successful interaction.

Mukawa, Oka, Arai & Yuasa (2005) put two subjects into interaction via coupled teleprompters and compared two conditions. In a first condition, the setup presents a direct frontal view of the partner, enabling mutual gaze. In a second condition, the cameras are

moved out of direct line of sight and film from the side. In this condition, no mutual gaze can be established. In both cases, subjects are asked to discuss their personal opinion on a topic that is imposed by the experimenter.

Mukawa et al. found that in the first condition subjects spontaneously accepted the functioning of the setup, get involved in face-to-face interaction and show ordinary behavior. They relate this acceptance to the possibility of successfully establishing mutual gaze. In the second condition in contrast (when subjects only see a side view of each other), subjects give and ask for numerous additional signals in order to confirm being in contact. With gestures and vocal confirmations, they try to compensate for the missing of mutual gaze, which is the usual signal of mutual attention.

As part of an experiment on the perception of the quality of an avatar in face-to-face interaction, Garau, Slater, Bee & Sasse (2001) realized a similar setup (see also following section). They use a teleprompter link, denominated in their experiments as ‘video tunnel’, as a reference for the comparison of different levels of gaze animation for avatars. This condition is of course rated better than audio interaction enhanced by different levels of avatar animation. Again, it is interesting to note that mediated interaction is maintained successfully without major difficulties.

Nowadays people are commonly acquainted with the visual presentation of humans on TV or computer screens. The interaction via devices such as the teleprompters or videophones is not very common and subjects may be sensitive to the peculiarities of such unfamiliar media. From the above-cited reports, we can however conclude that these interaction devices do not put the subjects into difficulty. They accept the setup and use it intuitively. Unfortunately, the above-described experiments did monitor neither measures nor observations that could be useful to acquire further knowledge about such mediated interaction between humans. This might have been of great interest for a comparison with our own experiments.



Figure 22: View of two person communicating via the ‘video tunnel’ (coupled teleprompters) as used by Garau et al. (2001).

1.4.5 Interaction between humans and animated agents

As mentioned above the findings by Argyle & Dean, Argyle & Cook and Kendon are frequently cited and used in works about gaze behavior in human interaction. Several researchers use their findings in the context of virtual animated characters.

Using these results, Garau et al. defined rules for the generation of gaze behavior of an avatar. In an experiment with 100 participants, they analyzed the impact of different levels of gaze animation on the quality of communication. Two subjects are put into mediated face-to-face interaction via teleprompters, denominated ‘video tunnel’ in their experiments. On the screen, the subjects are represented either by their video image, or by an avatar. The experiment

consisted of four different conditions of representation of subject: audio connection only, an avatar with randomly animated eye gaze and head orientation, an avatar with gaze animation driven by a model based on the mentioned literature or a direct video connection with accompanying audio signal. In the avatar conditions, the lip movements were animated based on analysis of the audio stream. The sex of the avatar is chosen to correspond to the sex of the subject it represents. In the condition with model-based gaze animation, the head orientation of the avatar was driven by motion capture of the subject's head.

Depending on the two modes 'speaking' and 'listening', the frequency, duration and direction of the animated agent's gaze are manipulated (see also section 1.6.3.2, page 64). Gaze may adopt two states: directed towards the partner or away. The position of the partner is not tracked and is assumed to be straight ahead from the middle of the screen. The gaze directions are chosen at random from a uniform distribution of angles between 0° and $\pm 15^\circ$ degrees away from the center.

Two different roles with a predefined initial situation in a time limited negotiation task are assigned to the interacting subjects. For the statistical evaluation of quality of conversation, different aspects such as impression of face-to-face, involvement, co-presence and appreciation of the partner are rated by the subjects in a questionnaire.

The analysis shows significant differences between the ratings of the different conditions. The video condition was rated best, followed by the condition using tracked head movements and gaze modeling for the avatar animation. The audio only and random gaze condition are rated worst and are not significantly different from each other. Garau et al. conclude that an avatar could enhance quality of telecommunication compared to pure audio, if it respects a certain coherence of behavior with the conversational flow, especially considering gaze control.

An experiment with a similar motivation (Garau, Slater, Vinayagamoorthy, Brogni, Steed & Sasse (2003)) but conducted in an immersive virtual environment confirms these conclusions. In this case no head tracking is used and a gaze model as described by Lee, Badler & Badler (2002) is used, that distinguishes speaking and listening states in a similar way as mentioned above. Two different qualities of realism of the appearance of the avatar are compared. The evaluation shows that independently of head-tracking gaze control can have a positive effect on the rating of interaction in an immersive environment if a minimum of realism of visual appearance of the avatar is provided.

Minato *et al.* (2005) conducted experiments to compare the gaze patterns of the human partner during human-human interaction and human-android interaction. They consider the measurement of gaze behavior as a method to evaluate human likeness of a humanoid robot. The gaze behavior of the subjects while thinking about the answer to a question asked either by a human or by an android interlocutor is evaluated. The robot is equipped with a highly human-like physical appearance. The gaze behavior is only characterized by percentages of gaze towards regions on the face repartitioned in a polar grid with the center at the bridge of the nose. Unfortunately, the repartition of gaze over target zones such as the eyes or the mouth cannot be reconstructed from these results. The experimental conditions distinguish between questions that needed thinking before answering and questions that could be directly answered, such as the date of birth. The latter type of questions should either be answered correctly or by a lie according to instructions given by the experimenter.

In general the questions requiring thinking result in less mutual gaze. This confirms the observations made by other researchers who found that people avoid gaze while thinking (Kendon (1967)). However, differences in gaze behavior are observed depending on the interlocutor. With a human interlocutor, the subjects gaze mainly downward when averting gaze. The background of Japanese culture of the subjects is proposed as a probable explanation. During interaction with the humanoid robot, averted gaze is distributed more equally over directions away from the straight ahead direction. When giving deceptive

answers the subjects keep gaze averted for longer periods. The gaze distributions are different when answering to a human compared to the android. Minato et al. conclude that there is a social component to gaze behavior. This confirms the observations by Gullberg & Holmqvist (2001) (see section 1.4.3.4, page 52).

The experiments on gaze animation of artificial show that humans do expect and need human-like gaze behavior for effective interaction with human-like agents.

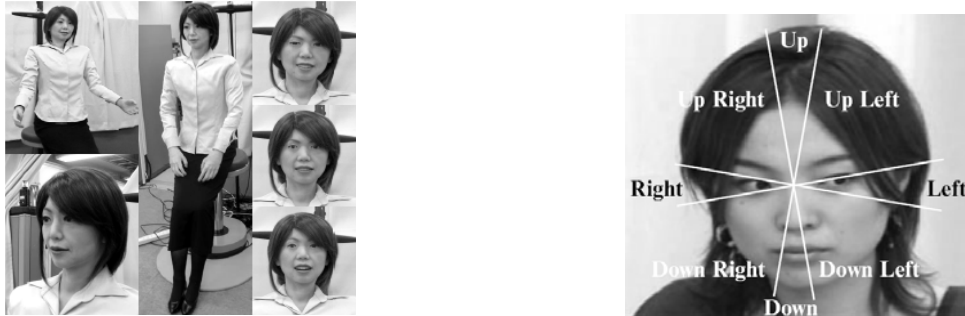


Figure 23: Left: Images of the robot used by Minato *et al.* (2005) in experiments on human-android interaction. Right: Repartition of regions of the face for the distinction of gaze directions.

1.5 GAZE AS INPUT MODALITY IN HUMAN-COMPUTER INTERACTION

The use of eye gaze as input modality for interacting with machines, aims at replacing the function of the computer mouse for object selection on a computer screen. Different strategies have been applied for this purpose. The simplest way is to define a threshold for fixation time to trigger the selection.

Jacob (1993) argues for the use of eye gaze for a non-command-based interface. He implements an interface using eye gaze in a way that deduces the user's interest from the scan path. The system preprocesses the raw gaze data measured with an eye tracking device. From the raw data, fixations can be detected, and when directed to the same target they are grouped into a single gaze. These fixations should be attributed to screen objects taking the knowledge about their distribution on the screen into account. This is a helpful approach to reduce the unwanted triggering of actions, which is known as the 'Midas-touch' problem in this domain. In the context of human-machine interaction, the mentioned grouping of fixations might be a useful strategy to detect user attention (see also Santella & DeCarlo (2004)).

In most applications using gaze as part of the multimodal input score in human-machine interaction, gaze is interpreted together with other modalities, considering also the dialogic context. Strategies that use gaze as unique input modality are not intuitive and require training. In a real physical environment, objects are not directly reactive to gaze nor speech. Only agents are sensitive to gaze. It seems thus more natural to use gaze to direct attention of artificial embodied agents and then to trigger appropriate actions.

1.6 EMBODIED CONVERSATIONAL AGENTS

The work reported in this document aims at giving gaze skills to a conversational agent. We thus give a brief overview on the origin of embodied conversational agents and techniques applied for their implementation.

The word 'agent' is first of all a term describing a recent programming paradigm. It can be seen as evolving from the previous paradigms of structural programming and object-oriented programming. Structural programming distinguishes data that is manipulated and the functions that operate on the data. Object oriented programming integrates the properties of

the data into the programming process, taking into account the relations between the data and the manipulations they will experience. The fundamental aspect of agent-oriented programming is the autonomy that the software gains through the consideration of the available resources (see <http://www.agentlab.de/aose.html> by Jürgen Lind). Different software agents are usually distinguished, as for instance multi-agent systems, autonomous agents, intelligent agents, distributed agents etc, according to their different applications.

‘Conversational agents’ are designed to communicate linguistically. Different modes of expression may be used, such as text or speech. When such an agent has a bodily representation, it is considered as an ‘embodied conversational agent’, commonly abbreviated with the term ‘ECA’. Dautenhahn, Ogden & Quick (2002) claim that a mere bodily appearance is not sufficient for an effective embodiment. The bodily appearance should in fact contribute to establish a relation of mutual influence between the agent and the environment. They argue that the extent to which the embodied agent and the environment may affect each other specifies the degree of actual embodiment. Embodied conversational agents typically use multiple modes of conversation exploiting the potential associated with the means of communication provided by their visual appearance.

The appearance of an embodied conversational agents can have a very basic form as for instance in the case of the agent ‘Eugene the Cuttlefish’ (Wallis (2005)). This agent dialogues exchanging written text. Its embodiment is restricted to a colored window on the screen (similar to HAL’s red light in Kubrick’s 2001 – A Space Odyssey). The color can be changed according to demands of the user directed to the agent or by the own initiative of the agent in order to signal basic emotions (such as fading to black when withdrawing from the conversation).

The appearance may also be basic in terms of body parts as in the case of cartoon-like characters. Gandalf for instance is a cartoon-like embodiment used by Thórisson (2002). His interaction model is a very comprehensive behavioral model, based on review of literature on psychological research as for instance the work by Kendon and Argyle & Cook (see also section 1.4.3). Bailenson *et al.* (2001) used a full body appearance but only animated the eyes. Buisine, Abrilian & Martin (2004) used modular drawings of full body cartoon characters that can be assembled in order to represent different gestures and attitudes but without realizing any movement.

More complex embodiment or more flexibility in the animation of the virtual character have also been proposed. Greta (Pelachaud (2002) Poggi, Pelachaud, de Rosis, Carofiglio & de Carolis (2005)) is a full body conversational agent with very detailed control of articulation, facial gestures and limbs that are jointly animated for conversation, interaction and emotional behavior. Likewise, REA (J. Cassell, Bickmore, Billinghurst, Campbell, Chang, Vilhjálmsón & Yan (1999)) has a fully articulated graphical body and is capable of speech with intonation, facial display, and gestural output.

An important difference between embodied conversational agents and animated characters as used by the film industry in animation films is the existence of a real-time model for animation. In order to meet the requirement to be interactive, the model underlying the generating of the ECA’s multimodal behavior must consider a minimum of input, such as for instance from scene analysis, speech recognition or simply text input. The model at the basis of the embodied conversational agent provides it with a minimum of autonomy in its actions and reactions to the scene as perceived through the means of scene perception. Animated characters may dispose of some physical model describing the degrees of freedom and the coordination of limbs during motion. The course of motion itself however has to be described by the animator or is copied from motion capture of human actors. There is no model that converts higher level targets into a motion sequence. Animated characters need explicit

control at any instance of their animation. This offers however the advantage of unrestricted flexibility of appearance.

Graphical representations of characters that are at least partly controlled by a human agent like in a magician of Oz scenario are denominated as ‘avatars’ (Bailenson *et al.* (2001)). The control parameters may come from motion capture, be derived from speech decoding or from command inputs. They are translated into commands animating a graphical representation. This procedure can be accomplished at different levels of modeling. A common application of avatars is the representation of a human participant in virtual immersive environments. Garau *et al.* (2003) for instance used an avatar as a graphical representation of interlocutors with articulation and head movement driven from the speech signal of a human interlocutor. Avatars offer a valuable possibility to separate higher level modeling from low level motor control and graphical representation for experimental evaluation.

1.6.1 Talking heads

Talking heads are 3D animated graphical representations of human-like heads or faces, usually with the aim to visualize speech through animation of facial features. Such a talking head is an ideal basis for the development of an embodied conversational agent. It should provide a sufficient number of degrees of freedom for its animation to be able to enrich speech production with non-verbal communicative gestures. We claim that the implementation of coherent gaze animation is crucial for the embodiment of such an agent.

A great variety of talking heads has been developed with different objectives and expectations concerning knowledge to be acquired. The most detailed approaches intend to replicate the facial muscles and to isolate their respective contribution to articulation and other facial gestures. Other approaches try to isolate modifications of the shape of the face and head that contribute to articulation without analyzing the underlying muscular structure and physiological mechanisms.

Physiological approaches model the rigid structure of the bones like the skull and the jaw, the active muscle components and the passive contribution of viscoelastic tissues like the skin, as well as the passive contribution of the rigid structure and muscles when not actively contributing. The integration of these contributions and their interactions as well as the coordination of the various muscles involved in particular movements, is a highly complex problem. This results in a control space that is much more complex than the degrees of freedom of the actually animated facial geometric structure. Ekman & Friesen (1975) and Ekman & Friesen (1978) (cited in G. Bailly, Béjar, Elisei & Odisio (2003)) established the Facial Action Coding System (FACS) for the coding of muscle actions to control almost all facial expressions (46 action units).

Approaches that aim at modeling the shape of a head as the consequences of physiological mechanisms but not considering them by itself, describe the changes of geometry resulting from these deformations. In general, they use a mesh structure to describe the head as a 3D graphical object. The different elements of the head are specified at different levels of accuracy according to their respective complexity and the intended precision of modeling. To generate movement, the displacement of vertices of this mesh structure can be interpolated between target positions or calculated according to a trajectory formation system. They can be controlled individually or as a whole set of vertices in a defined and coordinated way, representing a complex articulatory gesture. For a realistic appearance, the mesh representation needs color rendering. This can be generated by texturing complemented with shadow effects modeled from the 3D structure. Well-known examples using the geometric approach are ‘Baldi’ from the Perceptual Science Lab at the University of California (Massaro, Cohen & Beskow (2000)), the characters developed at the Department of Speech Music and Hearing at KTH in Stockholm (J. Beskow (1995), J. Beskow, Dahlquist,

Granström, Lundeberg, Spens & Öhman (1997)) as well as virtual characters complying with the MPEG4 standard.

Another approach to generate talking heads uses image-processing techniques with the potential to generate a highly realistic visual appearance. VideoRewrite (Bregler, Covell & Slaney (1997) cited in G. Bailly *et al.* (2003)) selects cutouts of the mouth from an image corpus that are superposed to a background video, applying different strategies to smoothen the transitions. A similar technique with higher degree of decomposition of the face is applied in the ATT Talking Face by E. Cosatto & Graf (1997; Eric Cosatto & Graf (1998) (cited in G. Bailly *et al.* (2003)). This enables more flexible and independent control of the different parts of the face but complicates the blending process and is more susceptible to artifacts. 'MikeTalk', an image processing technique at pixel level is proposed by T. Ezzat & Poggio (1998) (cited in G. Bailly *et al.* (2003)) using optical flow. Using optical flow they calculate the displacement and color change of pixels between target images and interpolate them to generate missing pixels. A more general model of combined shape and appearance was introduced by Cootes, Edwards & Taylor (2001). Active Appearance Models (AAM) are now largely used for model-based tracking of moving and deformable objects including faces. Theobald, Bangham, Matthews & Cawley (2001) use also AAM for facial animation.

A common technique to generate fluent movements of the face is to interpolate between target representations. Similar to 'phonemes' in speech synthesis, these target images are called 'visemes'. This technique can be extended to the concatenation of whole sequences of video chunks (Fagel (2006), Weiss (2004)). The interpolation between target representations has to be controlled in order to generate an appropriate transition such as respecting coarticulation effects. Cohen & Massaro (1993) propose a technique using spatial and temporal weighting of parameters. Furthermore, the synchronization between visual targets and speech has to be managed as the timing between salient audiovisual events depends on coarticulation and regime of phonation.

The embodied conversational agent used for the experiments reported in the present thesis is based on the talking head developed at the Department of Speech and Cognition at GIPSA-Lab. This talking head uses a graphical, data driven approach. It is based on the facial movements of a real person. The face of a subject has been recorded while speaking and the deformations have been captured and analyzed at different levels of precision. Colored beads are glued on the face with a density depending on the expected complexity of deformations. The face is then recorded during various activities ranging from text reading (Badin, Bailly, Revéret, Baciú, Segebarth & Savariaux (2002)) to more spontaneous face-to-face conversation (Gérard Bailly, Elisei, Badin & Savariaux (2006a)). The displacement of the beads is analyzed in order to reduce the degrees of freedom inherent in their arrangement, which resulted in six basic parameters to control articulatory movements for speech. There may be correlations of deformation between distant parts of the face that may be subtle and not expected but still important for realistic appearance and correct perception. The six parameters explain usually more than 95% of the variance of facial deformations and predict facial geometry with accuracy close to the millimeter.

To complete the talking head for other facial gestures, further parts of the head have been analyzed. A detailed model for the eye and eyelids has been developed and movements of eyebrows and the neck have been added (Casari (2006)). Context dependent recordings have been made to be able to model changes in attitude that permit for example the generation of basic expressions such as smiling or disgust.

1.6.2 Animation of Embodied Conversational Agents and Modeling of interaction

The animation of conversational agents is a complex process that needs modeling at different levels of abstraction and complexity. An approach to solve this problem is for instance

SAIBA (Situation, Agent, Intention, Behavior, Animation) (Kopp, Krenn, Marsella, Marshall, Pelachaud, Pirker, Thorisson & Vilhjalmsson (2006)). It is a joint development project by independent research groups in the domain of embodied conversational agents. It is claimed to be independent of domain and application and independent of the model of realization of graphical and audio appearance. It aims at separating the information to be conveyed from the animation process. In order to define a universal framework for multimodal generation of behavior, the generation process is separated into three levels of abstraction. At the lowest level, the intent of a behavior is planned and a description of communicative and expressive intent is transmitted to the next level, which is responsible for the planning of behavior. From this level, multimodal descriptions of behavior are given to the realization level that is in charge of transforming these transcriptions into graphical or audio representations. SAIBA leaves the modules situated at these levels open to individual solutions but defines the representation and bidirectional exchange of information between these levels in XML compliant coding languages. This should facilitate modular architectures and the exchange of modules between developers.

Peters (2006) for instance implemented a model for the generation of interaction activity and strategies for virtual agents in a virtual environment. It implements a basic theory of mind mechanism (without realization of 'SAM') to generate hypothesis on another agent's visual perception and states of mutual attention based on the input from other modules. It is inspired by the description of human mental processes during interaction as a modular system proposed by Baron-Cohen (1995) (see section 1.4.2, page 45). From the results of visual scene analysis, inspired by the work of Itti *et al.* (2003) (see section 1.6.3.1, page 62), it detects the attention of another agent from its gaze direction, body posture and movement and checks for mutual attention notably detecting mutual gaze. Using memory of this scene analysis the theory of mind mechanism generates beliefs about the mental states of the other agent. From these beliefs and the own motivations of the agent a decision about initiation of interaction and appropriate strategies for action are undertaken.

An aspect of planning of interaction and behavior that is not addressed by the SAIBA approach are the constraints in reaction time for different types of behavior. If behavior needs planning at different levels of complexity, it also needs an analysis of the scene and analysis of the behavior of an interlocutor in order to generate appropriate reactions. This is a crucial aspect for the ability to establish a grounded interaction and to signal attention. Therefore, the expectations towards the different delays of reaction have to be respected in order to be believable. The generation of dialogue for example requires analysis, comprehension and synthesis of dialogic acts with time scales of the order of a few utterances. Reflex reactions require prompt reactions according to the results from the scene analysis, such as involved in eye contact or exogenous saccades. Thus, different loops of interaction with different processing rates emerge.

This need for difference in processing time is addressed by the model for turn taking developed by Thórisson (2002). It is based on psychological research on human behavior in face-to-face dialogue in western culture. The model is structured into three layers of priority (see Figure 24, right). They correspond to loops of different processing time and account for the different time scales of actions observed in face-to-face interaction. The fastest interaction loops of durations shorter than 500ms are treated by the 'reactive layer' and have the highest priority. They mainly concern gaze behavior such as reactions to exogenous events. Thórisson considers these reactive actions to have a lower need for accuracy, which at the same time benefits processing speed. The basic management of dialogue, such as the control of beginning and ending of turn, is treated on the process control layer. The duration of a loop cycle is typically of 500ms to 2s.

The actual matter of the discussion is treated in the ‘content layer’, which has the lowest priority and loop speed but very high accuracy in the decisions taken. This layer manages the speech production and multimodal behavior related to the topic of the dialogue.

The system promotes a parallel processing of data such as for instance the interpretation of multimodal input and the generation of multimodal output. The activation of processes depends on the states of the system as well as the role of the agent in the current interaction and the perception-action reaction time required for the current activity.

A very rich multimodal scene analysis provides input for the system and is used at different depth of treatment by the three layers. The lowest level of data analysis is used in the reactive layer and covers positions and orientations of body parts and eyes of the interlocutor as well as basic analysis of speech data for intonation and syntax. The highest level involves for instance discourse comprehension as well as evaluation of age and personality of the interlocutor. This evaluation requires a minimal duration of conversation to proceed. The states of the upper layers of course condition the reactions of the lower layers (see Figure 24, left) such as being more polite with infants or elderly people than with teenagers.

Thórisson reports successful interaction of a large number of persons with the system. However, there is no objective validation of the model comparing implicit measurements of performance.

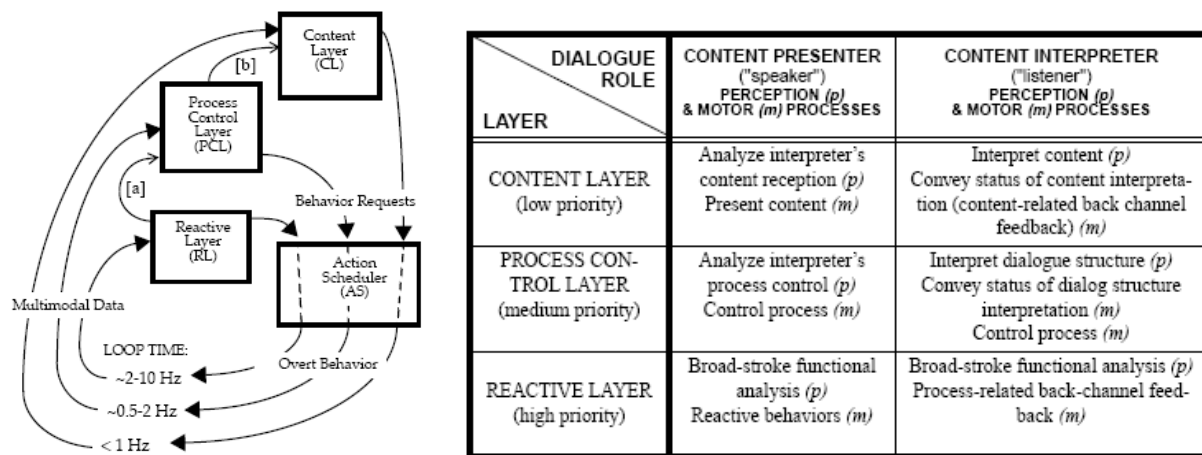


Figure 24: Left: Schematic representation of the model for turn-taking proposed in Thórisson (2002), separated in three layers of different processing time, that exchange data, analyse input from scene analysis and generate output for the generation of behavior. Right: Description of the three layers mentioned in the diagram to the left, detailing the level of priority and the type of in- and output data treated in dependence of the role taken in conversation.

1.6.3 Gaze Modeling for Conversational Agents

We claim that coherent gaze animation is a crucial factor for a believable presentation of embodied conversational agent. Several researchers addressed this problem and developed gaze animation models. Two different aspects may be distinguished in this context. There are models that address the animation of gaze in social interaction and its relation with conversational activity. Other approaches are dedicated to the modeling of gaze in relation with scene perception. This mainly concerns the emulation of reactions to exogenous stimuli that attract attention and trigger saccades or smooth pursuit. Some of these models of visual attention include top-down processes for proactive exploration of the visual space.

1.6.3.1 Modeling of gaze and scene perception

Itti *et al.* (2003) developed a model for visual attention and gaze movements inspired by the neurophysiology of visual processing of primates. It analyzes dynamic visual scenes by means of image processing techniques to determine where attention and gaze direction would be directed to.

They first apply a foveal filter that emulates the high resolution foveal perception in the central field of view and the progressive blurring of the scene away from it. As they use prerecorded video input without the possibility to align the camera with the determined gaze direction, this imitates changes of gaze and head orientation towards different items in the scene.

To analyze separately the different characteristics that may lead to visual salience, the frames are decomposed in different channels and their respective characteristics stored in different maps representing. Based on the hypothesis that similar features compete for saliency, whereas different modalities contribute independently to the perception of saliency, some of the features are combined into conspicuity maps. The features are determined at different resolutions to consider objects of different size separately.

The conspicuity map of color is calculated from complementary contrast of the colors red and green, and blue and yellow. An orientation map is calculated from the four maps of orientation in the main directions 0° , 45° , 90° and 135° . An intensity map is also calculated from the color channels. From differences between the current and previous frame, flicker is determined as the difference of absolute luminance and motion from spatial shifts.

These maps are weighted individually imitating neuronal processes of perception of contrast to isolate salient features and then added up. This is important to avoid that the final saliency map calculated from the individual feature maps is dominated by noise.

In order to add task-relevant constraints in the determination of saliency, certain regions can be attributed a higher priority in the saliency map to emphasize regions of increased relevance for the current task and to increase their probability of being attended to.

The maximum of the saliency map is chosen as the target for the next shift of attention, which is a covert shift of attention as it is not directly followed by a gaze shift. An overt shift of attention in the form of an actual gaze shift is only performed when four consecutive covert shifts of attention have been directed towards a same region at a minimum distance away from the current gaze direction. To avoid that a single item is kept as the target of attention, an inhibition of return mechanism has been added, that suppresses a previously attended target for a certain time.

Once a target for an overt shift of attention has been identified, the gaze shift is planned as a joint head and eye movement. Gaze shifts are realized as eye movement only when the shift is smaller than a threshold that depends on the initial position of the eye in the orbit. Otherwise, it is realized as a joint head and eye movement, whereas the gaze shift is treated separately for the vertical and horizontal components. The velocities for eye and head movement are functions of the amplitude of a shift. Moving objects can be followed with smooth pursuit.

During the execution of a saccade, the saliency map is inhibited which prevents shifts of attention and emulates the ballistic nature of saccades as well as the restricted visual perception during saccades. The time necessary to recharge the saliency map results in a delay of a succeeding saccade similar to the intersaccadic latency. A basic model to generate blinks takes into account the time passed since the last blinks or saccades as well as the magnitude of succeeding saccades to adjust the probability of a blink to appear.

The model by Itti *et al.* is used as a basis for several other implementations. Peters & O'Sullivan (2003) use a simplified version for bottom-up gaze generation for virtual agents from perception of the virtual scene.

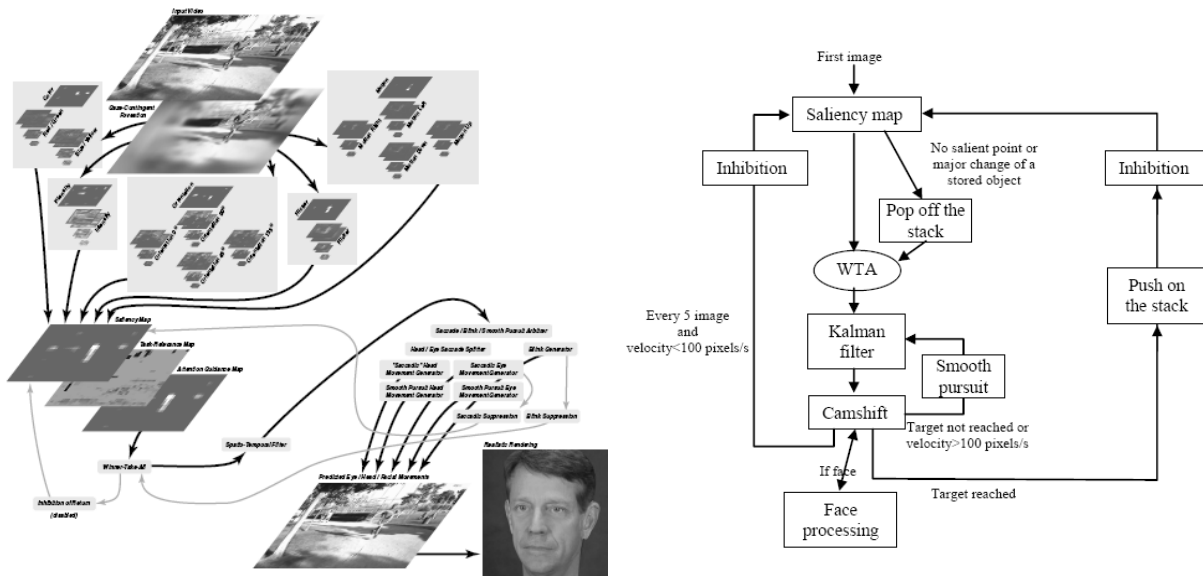


Figure 25: Left: Schematic description of the generation of gaze from saliency maps enhancing different features (motion, color, intensity, flicker, orientation, etc) of the video input used in the gaze model by Itti *et al.* (2003). Right: Chart of the model proposed in Picot, Bailly, Elisei & Raidt (2007) based on the computation of a similar saliency map but using an attention stack for switching between targets.

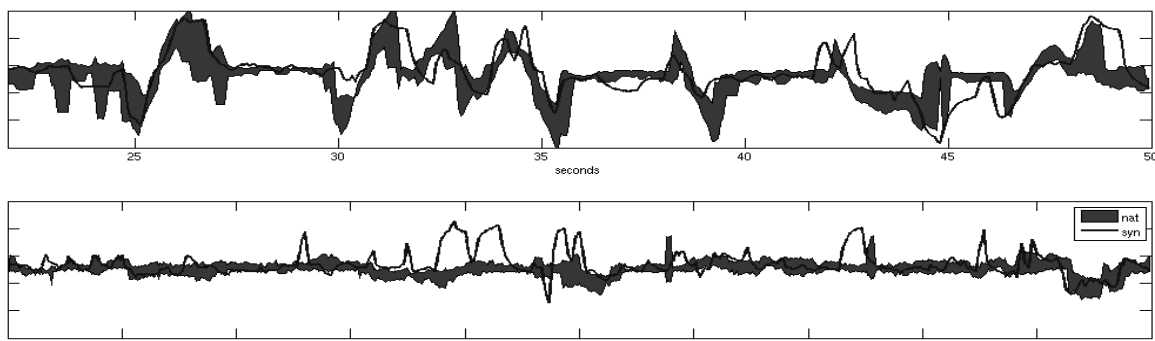


Figure 26: Comparing gaze direction as predicted by the gaze model proposed in Picot *et al.* (2007) (continuous black line), with gaze as recorded from subjects viewing the same stimulus (bold gray line). Top: right – left coordinates. Bottom: up – down coordinates.

Picot *et al.* (2007) reproduced and enhanced the model developed by Itti *et al.* (2003). They implemented a saliency map in a similar way from an orientation map of the two main directions 0° and 90° , a movement map, a color and an intensity map. Once a target of interest has been identified in an image as the most salient item, the corresponding graphical object in the scene is identified and its characteristics stored in a FIFO stack (first in first out). This is used to realize inhibition of return by subtracting the saliency of the objects stored in the stack from the saliency map when checking for new targets. When a new object is identified, the oldest object in the stack is discarded and can attract attention anew. If the number of objects does not exceed the stack size, the system loops over the objects in the stack until a new salient item appears. An object of extreme saliency, such as a fast moving object appearing in the scene, may interrupt the process. In this case, the stack is used as a LIFO stack (last in first out) to store the current object. The system attends to the new object and once this is treated gets back to the previous object, popping it off the stack. A Kalman filter is used to track moving objects. When the speed of a moving object exceeds a certain threshold this object is attributed with the highest saliency and therefore the attention maintained to it which produces smooth pursuit gaze. In order to take the importance of faces into account, a top

down approach for face recognition is added. Once a face is detected, the system tries to detect the eyes and the mouth as targets of priority on the face.

In an experiment, it was shown that the scan path predicted by this model is similar to the scan path of subjects viewing the same video stimulus (see Figure 26). This suggests that in most cases the model correctly attributed saliency.

1.6.3.2 Modeling of gaze and social interaction

A statistical model for gaze animation of a talking head in dyadic interaction was developed by Lee *et al.* (2002). It is based on empirical data on saccades and eye movement. It distinguishes talking and listening as two different modes of strong influence on gaze behavior. The statistics are taken from a single recording of eye gaze of a person with a head mounted eye-tracker during unrestricted dyadic interaction. The statistics reflect the dynamic characteristics of the observed eye movements, which include saccade magnitude, direction, duration, velocity, and inter-saccadic interval. However, it does not take into account the actual target of the observed gaze. The system is quite egocentric and always supposes that the interlocutor is facing the talking head.



Figure 27: Examples of the visual appearance of the talking head used in Lee *et al.* (2002). Left; straight ahead gaze. Right; averted gaze.

Garau *et al.* (2001) performed experiments to test the impact of avatars on the quality of communication and the differences between no gaze animation, random gaze animation, inferred gaze animation or direct video contact via video monitors (see section 1.4.5, page 54). The inferred gaze is derived from the audio stream distinguishing speaking and not speaking as two states influencing gaze. Possible directions are straight or averted gaze (see Figure 28). Again, the real location of the interlocutor is not taken into account. Averted gaze is directed randomly to targets at vertical and horizontal angles ranging uniformly distributed from 0° to $\pm 15^\circ$. The target and duration of fixations is chosen based on the observations reported by Argyle & Cook (1976) and Kendon (1967) (see section 1.4.3, page 47). Each pair of subjects was put into a negotiation task. A problem to discuss and a role as initial situation were imposed and left to individual interpretations. A mutually acceptable decision should be reached within ten minutes. Subjects filled out a questionnaire after the experiment and were interviewed by the experimenter.

The results show significant influence of role and gaze control on the estimation of quality of communication, with different impact in virtue of different aspect of quality. It can be concluded that the presence of an avatar can enhance the quality of remote communication but only if the animation is correctly related to the discourse.



Figure 28: Visual appearance of male and female avatars with straight and averted gaze, as used in Garau *et al.* (2001).

Based on the findings of Garau *et al.* (2001) and the proposed gaze animation model by Lee *et al.* (2002), Vinayagamoorthy, Garau, Steed & Slater (2004) and Garau *et al.* (2003) implemented another gaze model for an avatar in a shared immersive virtual environment. In the experiments, two subjects interacted, each represented by an avatar. The avatars had different levels of realism (see Figure 29). For the behavior model, they used talking, listening, looking at partner and looking away as internal states of the avatar. The speech activity is again detected from the audio stream of the interacting subjects. The ‘looking at partner’ gaze direction is oriented towards the head of the partner. The position of the partner’s head is monitored with a head tracking device. Together with the tracked position of the hands it is further used to animate the body of the avatar representing the subject. The ‘away’ gaze directions are chosen from eight possible angles spaced by 45° . The dependence of saccade velocity from the viewing angle was simplified compared to the model of Lee *et al.* (2002). It does however not take into account the animation of the surrounding of the eyes (lids, brows).

This model was compared to a random eye-gaze model that was built by reducing the inferred gaze model. In this case the angle dependent saccade velocities were set to a uniform velocity and the magnitude was taken from a uniform distribution of angles between 0 and $\pm 15^\circ$. The direction was chosen from a uniform distribution of angles between 0 and 360° . Independent of the factors listening and speaking the duration of fixation was kept constant at 2s for the random eye-gaze.

The scenario as well as the evaluation method is identical to those used in their previous study on quality of conversation enhanced by video presentation of a talking face with different quality of animation (Garau *et al.* (2001)). The results of their experiments that are relevant in the context of this thesis are that the more sophisticated the gaze model is, the more it is appreciated when used with the more realistic avatars. The simplified version is thus more appreciated when used with the cartoon-like avatar.



Figure 29: Avatars as used in Vinayagamoorthy *et al.* (2004) and Garau *et al.* (2003). Left: cartoon-like avatar. Middle: realistic avatars. Right: view of avatar in virtual environment.

Another model based on the work by Argyle & Cook and Kendon, is developed by Colburn, Cohen & Drucker (2000) for the gaze animation of an avatar. They defined two main states of eye gaze: 'looking away' or 'looking at' an interlocutor. The model does however not specify the 'away' direction and simply uses random values. The dwell times between transitions depend on the current state of eye gaze and the speaker. In the case of gaze directed at the interlocutor, the state of eye gaze of the interlocutor is also taken into account. If the latter is starting to look at the avatar while it is already in the 'looking at' state, this triggers a head nod of the avatar as a sort of acknowledgement signal of mutual gaze. The dwell time for this state is thereby extended by the duration of the head nod. For the transition between conversational states there are initial probabilities for the state of eye gaze to chose. The probability that the avatar will look at the interlocutor is set 'high' when it begins to speak and set to 'certain' when it begins to listen. The respective parameters are derived from literature. Colburn et al. extended this model to multi-party conversation with several avatars based on personal observations.

To test the gaze model, they conducted experiments involving subjects into conversations with the avatar representing the experimenter. The gaze direction of the subjects was monitored by a human observer. They found that using the gaze model instead of straight ahead gaze only, did not increase the mean percentage of time that subjects directed their gaze towards the avatar. A slight increase of gaze towards the avatar was observed while subjects listened to the avatar.

Heylen, Es, Nijholt & van Dijk (2005) implemented a virtual concert ticket shop consisting of a conversational agent behind a vendor desk. Subjects were asked to make reservations for concert tickets via this interface. The agent is equipped with a basic gaze model distinguishing three gaze directions. The decision of gaze direction follows the rule based model described by Justine Cassell *et al.* (1999) that aligns straight ahead gaze with the rhematic part of a sentence. Otherwise, the gaze is directed upwards into the air or at the documents on the table if this is suggested by the discourse. For comparison, a model choosing between the three gaze directions randomly and a reduced model that only distinguishes straight ahead and gaze towards the table have been implemented. For the analysis of user satisfaction, subjects had to fill out a questionnaire rating ease of use, satisfaction, involvement, personality and naturalness of head and eye movement. Furthermore, the processing time was measured. The results showed best performance and efficiency for the detailed gaze animation, which led the authors to conclude that already a basic gaze generation model can have a significant effect on performance and agreement.

Bilvi & Pelachaud (2003) implemented a model for gaze generation in dialogue. It takes text, tagged with labels of communicative functions, as input combined with a statistical model to generate eye movements alternating between direct and averted gaze. Note that no saccadic model is included and that alternation is paced by phone boundaries.

The evaluations of the different gaze animation models presented above show that they are appropriate to enhance the impact of embodied conversational agents. Subjects are able to differentiate between different levels of refinement of the models. These results are encouraging for the ambition to further investigate gaze behavior and improve gaze animation for embodied conversational agents. It might be interesting to develop strategies of evaluation that are apt to measure intrinsic parameters of the effect of gaze animation, such as cognitive load, credibility or engagement. Parameters like processing time as used in Heylen *et al.* (2005) are more appropriate to obtain unbiased measures and may unveil unconscious reactions to the system that are not be accessible with questionnaires.

1.6.4 Conversational Agents versus Robots

Animated characters and anthropomorphic robots are two platforms that may address the same topic. An important difference however is the bodily presence of robots implicit to their physical appearance. An embodied conversational agent needs strategies to generate the impression of presence and awareness of common environment for the establishment of grounding of interaction. For robots, this is not necessary to the same extent as already accomplished by their mere physical appearance. The requirement to signal attention and awareness is however common to both. The embodied conversational agent needs to establish and maintain mutual attention through interaction in both cases. Given the fact that it is able to localize and respect people, the robot needs to convince the users that its physical presence and activity are not dangerous for them. This requirement may be a reason why the robotics community developed sophisticated skills of scene analysis including analysis of human gestures and behavior, even if the development of a robot is not focused on human-like appearance and imitation of human behavior. These are skills such as the interpretation of pointing (Nickel & Stiefelhagen (2003)), acoustic speaker localization (Bechler, Schlosser & Kroschel (2004)) and extraction of facial features (Stiefelhagen, Yang & Waibel (1997)). In contrast, most ECA research is predominantly focused on rendering and animation.

Compared to robots, animated characters have the clear advantage of a development process that is more flexible but less expensive and time consuming. Furthermore, they have access to the real world, as well as to their virtual space of evolvment that they may use to present additional information.

Some of the multiple fields of interest of robotics are very well balanced concerning their skills of scene analysis and synthesis. Therefore, these are closer to the concerns of the ECA community and very interesting with respect to the present thesis.

At the Takanishi Laboratory at WASEDA University in Tokyo there are different research projects investigating aspects of anthropomorphic robots, such as bipedal walking, vocal tract modeling for articulatory speech synthesis or generation of different face shapes. The WE-4RII project is an upper torso and head robot that addresses questions that are also of interest for the research on ECA (<http://www.takanishi.mech.waseda.ac.jp/research/>). Its major objective is the expression of emotions in a human-like manner by the means of facial expressions, body posture as well as hand and arm movements. It emulates the human body in great detail, copying the articulations of the arms and hands and the facial movements such as lip and eye brow movement. Furthermore, it is equipped with a very rich scene analysis, imitating the human active visual (including head and eye movement) and auditory system, cutaneous sensations such as tactile and temperature sensation and even olfactory sensation by the means of gas sensors. The perception as well as the expression of the robot can be biased by definitions of its personality. The active behavior of the robot is motivated by a 'need model' generating internal stimuli. This includes appetite (energy consumption), need for security (withdrawal from strong stimuli), exploration (learning of relation between visual information and target property) and the need to show behavior. The robot changes its internal states of emotion in dependence of external and internal stimuli and generates according expressions (see Figure 30).

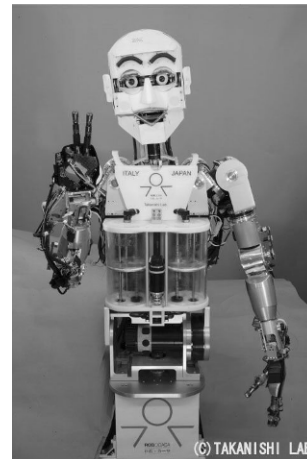
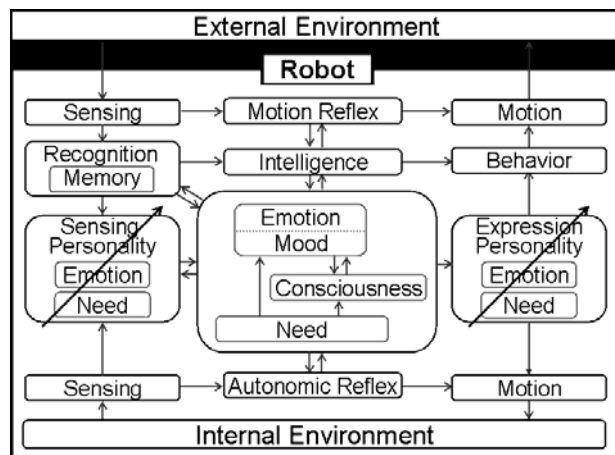


Figure 30: Left: Function chart describing the behavior model of the WE-4RII robot from the Takanishi Laboratory at WASEDA University. Right: Photo of the WE-4RII robot.

The Perceptual Computing Group at the Humanoid Robotics Institute of WASEDA University developed ROBITA as a robot multimodal interaction with humans (Matsusaka, Fujie & Kobayashi (2001)). It disposes of speech synthesis and recognition, face recognition and face direction detection. Furthermore, it can interpret pointing gestures from image processing and can equally execute pointing gestures itself. In a group conversation, it is able to signal its attention with body and eye movements directed towards the current or expected speaker. A module for turn taking determines who will take the turn when addressed to, or whether it should interrupt a discussion if the system detects that wrong information is exchanged. The interaction process uses the approach of interaction loops of different update frequencies. It distinguishes for example between low frequency update of speech information and the high frequency treatment of visual feedback, similar to the model proposed by Thórisson (2002).

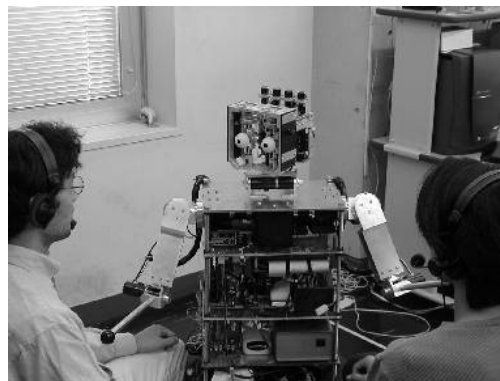


Figure 31: Photo of ROBITA, developed at the Perceptual Computing Group at the Humanoid Robotics Institute of WASEDA University, in interaction with two persons (Matsusaka *et al.* (2001)).

Another example of such an anthropomorphic robot is Kismet that has been developed by the Humanoid Robotics Group at the Artificial Intelligence Laboratory at MIT (see Figure 32, left). It is designed as an autonomous robot for social human-robot interaction. It takes the role of the infant in a caretaker-infant dyad (Breazeal (1998)). Similar to the need model of the WE-4RII robot, Kismet has drives. The drives are processes that seek to maintain their activation energy in a predefined range. This activation energy can be influenced by external events or by internal stimulation, implemented as a cyclic process. Furthermore, the behavior

of the robot has a regulatory influence on the activation energy. The robot produces behavior in order to satiate its drives.

By the means of facial gestures and body posture Kismet is able to display emotion that reflects the state of its drives. The processes are modeled as transducers that calculate activation energy from several weighted and biased input signals. Once the activation energy for a process exceeds a threshold, the process is activated. This may result in the display of behavior, the transmission of messages or the discharge of activation energy to other processes. The activation energy may be positive as an excitatory or negative as an inhibitory influence. The behavior and display of emotion are realized by the motor system that transforms them into corresponding motion.

Kismet is equipped with an active visual system imitating human vision (Scassellati (2001)) that executes movements such as saccades towards detected targets of interest and smooth pursuit (see Figure 32, right). It uses a bottom-up approach that is similar to the one used by Itti *et al.* (2003) to detect salient items (see section 1.6.3.1, page 62). In addition to the analysis of visual input for saliency of color features, it uses a trained detector specialized on skin color. The drives of Kismet have a top-down influence on the weighting of visual features, similar to the model proposed by Picot *et al.* (2007) (see section 1.6.3.1, page 62). When Kismet is seeking for human company, for instance skin color features are weighted as especially salient. Once skin color is detected, the system seeks to identify facial features to estimate head posture and to localize the eyes. The behavioral system of Kismet is highly inspired by the work of Baron-Cohen (1995) (see section 1.4.2, page 45),

A very interesting aspect of the display of emotion by Kismet, is that it corresponds to an actual state of the system and therefore carries meaning in respect to ongoing interaction. It may for instance signal that the system is overloaded and unable to treat the incoming data by retracting from personal space of interaction, or that it needs input stimuli when showing boredom. This is an appropriate technique to signal inner states of the robot that are of importance for the ongoing interaction. The displayed emotions provoke the interacting subject to react in response to the needs of the robots in an intuitive way, just as would be expected in human interaction. The design of the behavioral system as a basis for a robot able to learn from interaction, is impressively successful to contribute to the maintenance of interest and interaction generating the impression of joint attention.

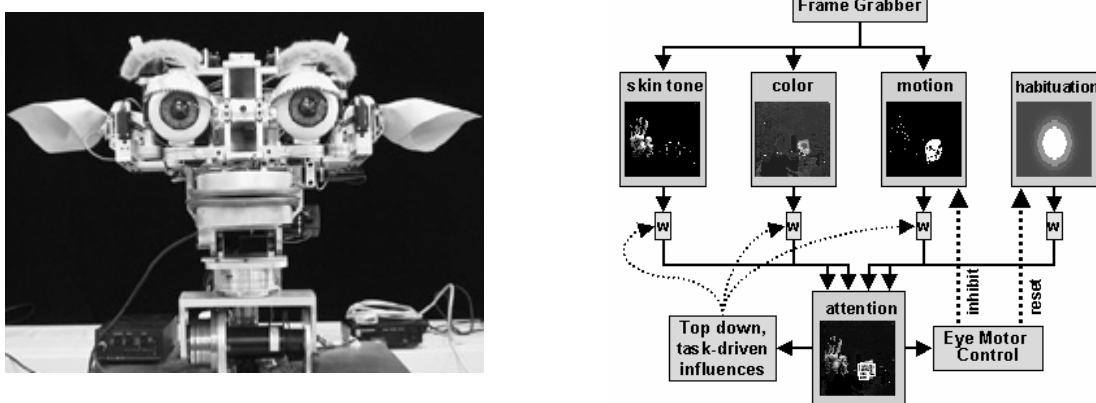


Figure 32: Left: photo of Kismet. Right: Schematic description of the active vision system implemented in Kismet – analysis of video separated in different channels; task-driven weighting of contribution from different channels for the determination of the target of attention; activation of motor system for gaze shifts towards targets of attention.

1.6.5 Evaluation and Perception of Conversational Agents

As conversational agents are designed for the interaction with human users, it is interesting to know how they are perceived by human subjects. There are different strategies to test the impact of features of an agent or to evaluate the agent as a whole.

A criterion to evaluate an agent, comparing it to a human as a reference, is known as the Turing test. It is named after Alan Turing who proposed a scenario where a subject should interact with another subject and a computer system at the same time (Turing (1950)). From the output, which for both was in the form of written text, the subject should decide which was the machine and which the human. This principle idea can be extended to systems that dispose of other modalities of interaction than written text only.

The scenario proposed by Turing is used to evaluate dialogue systems on the basis of written text for the annual awarding of the ‘Loebner Prize’. Geiger, Ezzat & Poggio (2003) extended this paradigm to audiovisual speech synthesis to evaluate the videorealistic talking head ‘Mary’ (Tony Ezzat, Geiger & Poggio (2002)). Subjects had to judge if a video was natural or manipulate. The audio signal was original in both cases. If the subjects were not able to separate correctly natural and synthetic animations, intelligibility provided by synthetic animation was however significantly lower. The subjects were thus unconsciously aware of the difference.

Usually only certain features are tested to investigate the impact of a new module and to compare it to an older version or the absence of such a module. Therefore, subjects are usually put into interaction with an agent that is endowed or not with the feature to test (see section 1.6.3.2, page 64 for tests of gaze animation). A common strategy of evaluation is the use of a questionnaire investigating the subjective perception of the agent and its behavior that are subsequently compared between the different implementations. A probably less biased evaluation can be made measuring objective parameters implicit to the interaction, such as reaction time, success or length of maintained interaction. Massaro *et al.* (2000) tested the intelligibility of his talking head Baldi, monitoring the successful recognition of utterances. Munhall, Jones, Callan, Kuratate & Vatikiotis-Bateson (2004) investigated the impact of head movement on the understanding of Japanese. In this study articulation and head movements were taken from the recording of a human speaker and imitated with a synthetic animated head. Thereby the head movements could be altered and their impact on intelligibility be compared independent of articulation.

In the case of gaze animation, the same strategies for evaluation may be applied. Experiments on gaze animation for conversational agents have shown that gaze animation that is related to the conversational activity of the agent improves the quality of the agent (see section 1.6.3.2).

A general problem in the context of evaluation of embodied conversational agents is a problem known as the ‘uncanny valley’. It describes the effect of a drop in the subjective evaluation of an agent when it approaches a very human-like appearance (Minato *et al.* (2005)). The human likeness is first perceived positively but leads to bad evaluation when subjects detect a discrepancy between the high degree of human likeness of the visual appearance and a comparably deceptive behavior due to shortcomings of its animation.

1.6.6 Conclusions

Speakers are in fact very sensitive to gaze behavior of their human or artificial interlocutors. Gaze should be sensitive to both exogenous stimuli in the scene and endogenous processes that should reflect well-known social behavior such as for instance being attentive to facial movements when listening to a person. This evaluation of gaze behavior of others monitors attention and influences exchange of information. Our challenge is thus to rely on precise

measurement of gaze patterns during natural face-to-face interaction and to reveal relevant factors of the observed behavior.

2 GAZE DIRECTION AND ATTENTION

We have discussed the basics of embodied conversational agents in chapter 1, where we have also shown in detail the importance of gaze in human interaction. Eye gaze was shown to be an efficient component of deixis in human interaction. For the completeness of its functionality and in order to establish grounded interaction with a human partner, an embodied conversational agent should incorporate such abilities. We intended to test to what extent the talking head developed at the ICP is able to generate persuasive deictic gestures, that provoke intuitive reactions of a user and how a user could benefit from these gestures. We particularly study here the impact of facial deictic gestures of the talking head on user performance in simple search and retrieval tasks.

2.1 THE EXPERIMENTAL SETUP AND SCENARIO OF INTERACTION

To set up an appropriate experiment, we followed up the experiments conducted by Langton *et al.* (2000) and Driver, Davis, Riccardelli, Kidd, Maxwell & Baron-Cohen (1999a) that built on the Posner paradigm. We reproduce the basic idea of a human head at the centre of the screen, the appearance of which influences the attention of a user, directing it to different regions on the screen. In contrast to these experiments, the head should be a permanent item on the screen and of central importance to the scene and the task. We chose an on-screen card game as a search and retrieval task to test the ability of the video-realistic face and eyes of our talking head to direct user attention in a complex virtual scene.

2.1.1 Setup

For the experiment we use a two screen setup with an ordinary computer screen for the experimenter and a Tobii® eye tracker screen as display for the subjects. The eye tracker monitors the gaze direction of the subject that can also be used as input modality. As further input modalities speech recognition, the computer mouse and the key board may be used. In the current experiment the input modalities are limited to the mouse.

The experiment is realized with an in-house multimedia scenario scripting architecture. It consists of an event-based language and a corresponding C++ code generator. This language allows the description and modification of multimedia scenarios. A finite-state automaton (FSA) describes each scenario as a series of states with pre-conditions and post-actions. As the user progresses in the scenario, the FSA specifies which states are waiting for events. Pre-conditions consist in conjunctions or successions of expected multimodal events. Such events are for instance recognized keywords, mouse clicks or displacements, eye movements or gaze towards active objects.

Pre-conditions can include testing of intervals between time-stamps which allows, for example, speech items to be associated with a certain focus of attention. Post-actions typically concern the generation of multimodal events. Time-stamps of these events can be used to delay their actual instantiation in the future. Post-actions can also generate phantom events that will be considered as potential triggers of pre-conditions of the states of the FSA. They can simulate multimodal input or function as a method to share information

Each input device emits events according to user actions associated with the internal model of the space of interaction that is refreshed constantly (see Figure 33:). Areas on the screen can be defined and be put under surveillance as selectable icons for the triggering of actions. The system posts new events such as entering of zone, quitting of zone and further parameters

such as for instance duration of fixation . The FSA is called each time an event is generated or updated.

For accurate recording of the timing and involved events, C++ source code is generated from the script and then compiled into a binary executable. The benefits of using C++ such as the use of variables, procedural and complex algorithms, remain accessible through code inclusion inside any script.

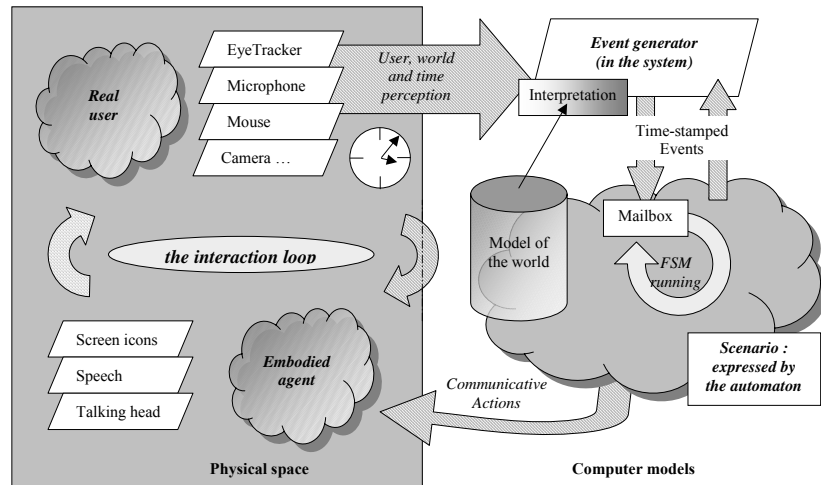


Figure 33: Schematic description of the finite state machine controlling the interaction of the experimental setup used for the card game experiments.

2.1.2 Scenario

As scenario we chose a card game as central task for the users, instructed to search a given object in a complex scene displayed on a computer screen. To either side of the screen four numbered and colored cards are displayed (see Figure 34). Another card, the play card, is displayed at the bottom of the screen at the beginning of every cycle. The talking head is displayed in the center of the screen. To start a cycle, the user has to click on the play card. As a consequence previously invisible digits will appear on the cards. The task consists of putting down the play card on the card with the same digit. If a wrong position is selected to put down the play card on, a digit on the card is replaced by an 'X' to indicate the choice as invalid. To anticipate memory effects the numbers on the cards are shuffled before each turn. The target position is alternated randomly, but uniformly distributed amongst the eight possibilities provided that the number of cycles is a multiple of eight. This should compensate for possible influences of the target position on the user performance. The color of the cards is associated with the position and is not changed during the experiment. The digits displayed on the cards are assigned at random, independent of the position and color of the target card.

For the investigation and detailed understanding of the impact of the indications given by the talking head, we realized different conditions of the experimental scenario. They correspond to different levels of assistance and help. The condition *'no assistance'* does not display the talking head and only the white background is visible at the center of the screen. In this case the subjects have to search the target card amongst the eight possibilities without any assistance, by comparing the play card to the digits on the other cards. In the condition *'correct indications'*, positive assistance is given by the talking head, indicating the position of the target card with deictic eye and head movements. Therefore an eye saccade is executed combined with a head turn into the direction of the target location. A variation of this condition is the condition *'no digits displayed'*, where positive assistance is given by the talking head, but no digits are displayed on the cards. The target card can only be found from interpretation of the deictic head and eye gestures. Finally, the condition *'deceptive indications'* with misleading indications was realized, where the talking head indicates at random a false card that does not show the sought-after digit.

In this case, the subjects have to find the target card by comparing the digits themselves despite of the indications of the talking head. Subjects are informed about these different conditions beforehand.

Figure 34 shows screenshots of the game interface with examples of the different conditions of the experimental scenario.

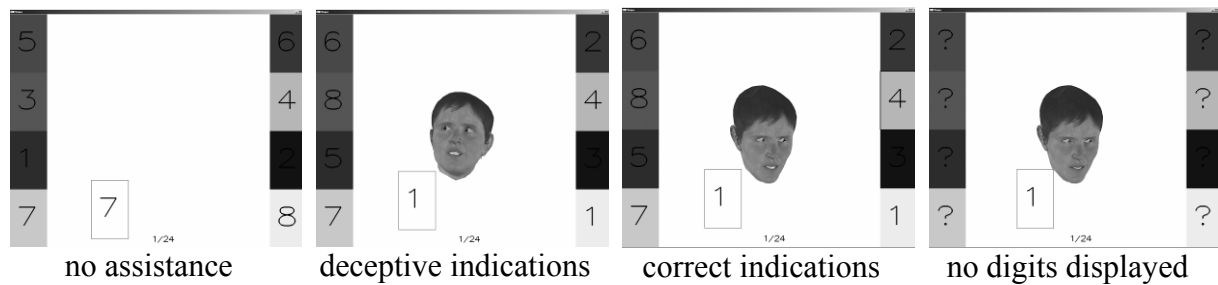


Figure 34: Experimental conditions: The experiment is divided into four conditions with different levels of help and guidance by the talking head. When the talking head is present it can give helpful or misleading facial cues. When the numbers on the cards are not shown (right; no digits displayed) these cues are the only possibility to locate the correct target card.

An experiment comprises four blocks of cycles of the same condition that are further denominated as ‘experimental conditions’. Each block begins with three training cycles to allow the subjects to become familiar with the condition and is then followed by 24 measurement cycles. The characteristics of the upcoming condition are announced as text on the screen before the training cycles begin and thus inform the user what behavior of the clone to expect. No advice how to interpret the presence and activity of the talking head is given. The user is instructed to put down the playing card on the target position as fast as possible but no strategy is suggested. General information explaining the task is given as text on the screen at the beginning of the experiment. Users were not informed about the data to be measured. An experiment took the time of about 15 minutes per subject.

We performed two series of experiments with slightly different conditions:

1.) Experiment I:

Experiment I is arranged as fixed sequences of the conditions *no assistance*, *deceptive indications*, *correct indications* and *no digits displayed*, as shown in Figure 34. We expected to obtain a neutral reference observation when there was *no assistance*, slower and less precise reactions with *deceptive indications*, and fast and correct responses with *correct indications*. The condition *no digits displayed* should reveal the spatial precision with which the deictic gestures can be interpreted. When the talking head was present, it also uttered sentences of felicitation and motivation in order to motivate the subjects after every successfully accomplished cycle. The utterances were generated off-line to avoid computation delays.

2.) Experiment II:

It appeared that during the first series of experiments the utterances of felicitation and reward were too repetitive, unprofitable and tended to irritate and annoy the subjects. They have therefore been suppressed for the second experiment. In experiment II the deictic gestures were accompanied by the short utterance ‘là’ (English: ‘there’). This transforms the previously purely visual gesture into an audio-visual gesture that does not only specify the location of interest, but also the timing of the attention shift. This is closer to natural behavior and as a short and clear instruction more appropriate to the demand of fast reactions. With this additional utterance, the gesture is expected to be more successful to attract and direct the attention as it addresses an additional sense of perception in a coherent way.

2.1.3 Data acquisition

For the evaluation of the experiments we perform an objective estimation of user performance as well as a subjective evaluation of the usability and acceptance of the talking head. As objective measure reaction time and gaze behavior have been monitored. The measurement of reaction time was started by the mouse click on the play card that triggers the cycle and reveals the digits on the cards. The click that puts the play card down onto the correct target card terminates the measurement. In order to be able to monitor the gaze behavior of the subjects a screen with an embedded eye tracker¹ was used as display for the card game. For a correct measurement a short calibration of the eye tracker has to be performed with every subject before the recording. The eye tracker computes the currently fixated point on the screen and records the data separately for the two eyes. From this data we compute which objects on the screen had been inspected and how much time users spent on them. We survey the eight target cards on the sides of the screen and the face of the talking head as possible fixation targets. Eye gaze towards the playing card is not monitored, as the play card is constantly moving during the experiment and smooth pursuit gaze would be too laborious to survey.

In order to obtain subjective evaluations, participants were asked to answer a questionnaire at the end of the experiment. They had to rank various aspects of usability and acceptance of the experimental conditions on a five-point MOS scale, and to choose which condition they considered as most appropriate and fastest.

2.1.4 Data Processing

The experiment was conducted on a two screen setup using the Tobii® eye tracker screen together with an ordinary computer screen. The computer mouse was active on both screens and could be moved from one to the other. Therefore, when trying to select a target card on the side representing the boarder to the second screen, overshoots were possible. These usually resulted in very long reaction times. A criterion was defined to detect such extreme outliers. Values of reaction time that are further than five times the inter quartile range away from the median value of the observed distribution made for a subject during an experimental condition, are defined as extreme outliers and replaced by the mean value calculated from the remaining valid data.

Considering the number of cards inspected during a cycle, which is calculated from the data measured with the eye tracker, a criterion was also defined to exclude invalid data. A cycle was only accepted as valid, if at least 95 % of processing time could be reliably surveyed with the eye-tracker. This concerns the time span between the mouse click to take up the play card, and the click to put it down on the correct target position. Invalid cycles may result for instance from a displacement of the subject's rest position out of the working space that is surveyed by the eye tracker. In these cases the eye tracker cannot reliably measure the gaze direction of the subject. As in some conditions numerous invalid cycles appeared, their replacement by a mean value, analogous to the treatment of reaction times outliers, would probably have altered the distributions of observations. Therefore invalid measurements of number of inspected cards were excluded from further analysis. To assure a sufficient number of observations for the statistical analysis, an experimental condition was only taken into account for further evaluation if a minimum of 60 % of the cycles were accepted as containing valid data, which corresponds to a number of 14 valid cycles. Otherwise the data was discarded from further analysis.

¹ Tobii® 1750 Eye-tracker

The corrected data was then analyzed statistically for differences of the means of the distributions observed during the different experimental conditions, subject by subject. For this purpose the values were first transformed to the logarithm of base ten in order to approach the distribution of observed values to a Gaussian distribution. Then ANOVA were performed for a pair-wise comparison of the conditions at $\alpha = 0.05$ separated for subject and separately for the variables reaction time and number of inspected cards. Only the combinations of conditions that are of interest to this investigation are discussed. If all cycles produced valid data there are 24 values per condition as factor of the ANOVA.

2.1.5 Variables of the analysis

In the experiments dedicated to the analysis of the impact of deictic gestures of a talking head, we conducted two separate series of experiments with different subjects. In both, we distinguish subject and experimental condition as independent variables. In both cases reaction time and number of inspected cards are measured as dependent variables as well as the answers to the questionnaire.

2.2 EXPERIMENT I: IMPACT OF FACIAL CUES

Ten subjects (six male and four female) took part in this experiment. They were researchers or students and ranged from 23 to 33 years of age. All regularly use a computer mouse and none reported vision problems. The dominant eye is the right eye for 9 subjects and the left eye for one subject. When controlling the data for extreme outliers and validity, 11 of 240 measured reaction times had to be corrected. This corresponds to 4.6% of the measured values. Never more than two values had to be corrected per experimental condition and subject.

2.2.1 Experimental conditions

During the first series, the experiment was arranged as fixed sequences of the conditions *no assistance*, *deceptive indications*, *correct indications* and *no digits displayed*, as shown in Figure 34. We expected to obtain a neutral reference observation when there was *no assistance*, slower and less precise reactions with *deceptive indications*, and fast and correct responses with *correct indications*. The condition *no digits displayed* should reveal the precision at which the indications can be interpreted. When the talking head was present, it uttered sentences of felicitation and motivation in order to motivate the subjects after every successfully accomplished cycle. The utterances were generated off-line to avoid computation delays.

2.2.2 Results

During the conditions with the digits on the cards visible, which permitted visual comparison of the playing and target card, only one wrong selection occurred. For these conditions, the task can therefore be considered as accomplished successfully. Numerous errors occurred (15% errors) during the condition where users had to rely on the gestures of the talking head when no digits were displayed on the cards. This indicates that a precise interpretation of the gaze direction of the talking head is difficult (see also discussion in section 2.4). Nevertheless, as all of these errors occurred between neighboring cards, we consider the assistance given by the facial gestures as helpful. When the user has additional information allowing visual comparison of the digits, as is the case during the other conditions, the deictic information seems to be sufficiently precise.

Pair wise comparisons of means of the distributions observed during the different conditions are displayed in Figure 35 for the reaction time and in Figure 36 for the number of cards in

the visual search path. Every subject is represented by a number corresponding to the order of participation. In the diagram, the order of subjects is sorted for increasing difference of reaction time means between the compared conditions from left to right. The order is kept the same for the figure displaying the mean number of inspected cards (Figure 36). When the pair-wise comparison with an ANOVA indicated that distributions are significantly different, this is indicated with a star above the number representing the subject in the diagram. The stars indicate significance of differences between the distributions of observations of the respective subject at $\alpha = 0.05$. A detailed listing of results from the ANOVA as F and p values is given in Table 1 and Table 2.

The representation of reaction time means in Figure 35 (left) shows that 5 out of 10 subjects had significantly shorter reaction times during the condition *correct indications* compared to the condition *deceptive indications*. These users have a substantial mean gain of reaction time of about 300 milliseconds per cycle. The mean reaction time over all subjects and condition is about 1.7 seconds. Comparing *correct indications* to the condition *no assistance* (second from left), this was the case for three subjects, with a mean difference of reaction time of about 400ms. Comparison of the condition *no assistance* to *deceptive indications* (second from right) shows in fact that these two conditions produce similar results. Only two out of 10 subjects show significant differences of the means that are contradictory in tendency. In the first case, subject 5 has significant shorter reaction times when no talking head is present whereas in the case of subject 9 this relation is inverted. During the condition *no digits displayed*, half of the subjects show longer reaction times compared to *correct indications* (right), which is probably due to the repeatedly occurring selection errors. In three cases subjects showed significantly shorter reaction times. In these cases this gain may be explained by a learning effect of interpreting the gestures of the talking head.

● no assistance; + deceptive indications; □ correct indications; Δ no digits displayed;
* $p \leq 0.05$

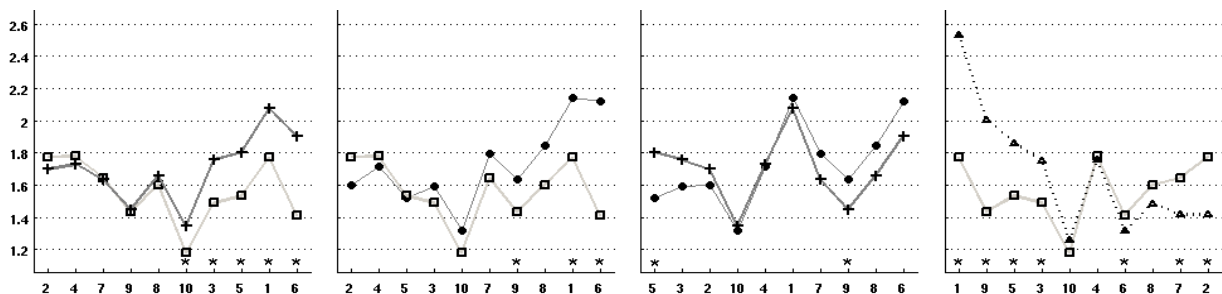


Figure 35: Comparing reaction time means for four pairs of conditions. From left to right: *deceptive indications* vs. *correct indications*; *no assistance* vs. *correct indications*; *no assistance* vs. *deceptive indications*; *no digits displayed* vs. *correct indications*. The x-coordinate lists the subjects whereas the digit represents the order of participation. Stars above these digits indicate statistical significance of difference between the underlying distributions at $\alpha = 0.05$. The order of subjects is sorted for increasing difference of reaction time means between the compared conditions from left to right.

Table 1: Information about age, sex and dominant eye (left or right) of subjects in order of their participation, as well as F and p -values as results of the ANOVA comparing pair-wise between conditions the distributions of reaction time. Results that indicate significance are printed in bold letters.

Subject	age	sex	dom. eye	correct ind. vs. deceptive ind.		correct ind. vs. no assistance		no assistance vs. deceptive ind.		correct ind. vs. no digits displayed	
				F	p	F	p	F	p	F	p
1	23	f	r	7.4	0.009	12.4	0.001	0.3	0.562	37.9	< 0.01
2	25	m	r	0.1	0.811	1.3	0.262	1.2	0.271	7.5	0.009
3	24	f	r	4.5	0.039	1.7	0.201	1.3	0.266	7.8	0.008
4	25	m	r	0.1	0.802	0.2	0.621	0.1	0.761	0.0	0.984
5	24	m	r	5.5	0.023	0.0	0.938	6.7	0.013	7.6	0.008
6	25	m	r	31.7	< 0.01	46.4	< 0.01	2.2	0.147	5.5	0.023
7	26	m	r	0.0	0.891	2.5	0.123	2.9	0.093	7.0	0.011
8	30	f	l	0.0	0.904	3.4	0.073	2.2	0.148	1.4	0.250
9	33	f	r	0.1	0.717	4.6	0.037	4.3	0.044	13.0	0.001
10	25	m	r	6.3	0.016	2.8	0.102	0.3	0.587	0.1	0.759

The evaluation of the data acquired with the eye tracker is shown in Figure 36. Some of the data was not sufficiently reliable and had to be excluded from evaluation as too often the gaze direction could not be monitored reliably. These are marked with ‘X’ above the subject number on the x-axes. Subject 7 is concerned in the conditions *correct indications* and *deceptive indications*, and therefore had to be excluded from all pair-wise comparisons. For subject 8 this is the case in the condition *no digits displayed* and the concerned pair-wise comparison. For the remaining comparisons sufficient reliable data was available. Over all subjects and condition we determined a mean of 2.5 cards inspected while searching for the correct target position.

Comparison of the means with an ANOVA at $\alpha = 0.05$ evidences an advantage for some of the subjects when correct hints are given by the talking head. The concerned subjects that profit from *correct indications* inspect in mean 1.5 cards less compared to *no assistance*. Those who profit from *correct indications* compared to *deceptive indications* inspect in mean 2 cards less. We interpret this as an advantage in terms of cognitive load since less cognitive resources are necessary for inspecting cards during the search process.

The topics treated in the questionnaire and the mean rating by the subjects on a five point MOS scale are given in Table 3 (see section 5.1 for exact wording of questions). From the examination of the ratings it can be retained that 5 of the 10 subjects think they are faster with the helpful assistance of the talking head and prefer it to accomplish the task. One of these even prefers to completely rely on the indications and not to have the option of visual verification. The fact that half of the subjects preferred not to have any assistance, suggests that the talking head may increase the cognitive load of the subjects. Also the quality of the talking head is not rated very high. During the condition *deceptive indications*, subjects suppose mainly that they do not pay attention to the talking head. The comparison of subjective answers of the subjects to the objective measure of their performance reveals no systematic relation between the two. In some cases they correspond, in others they are contradictory. It seems not possible for the subjects to reliably estimate their own performance correctly.

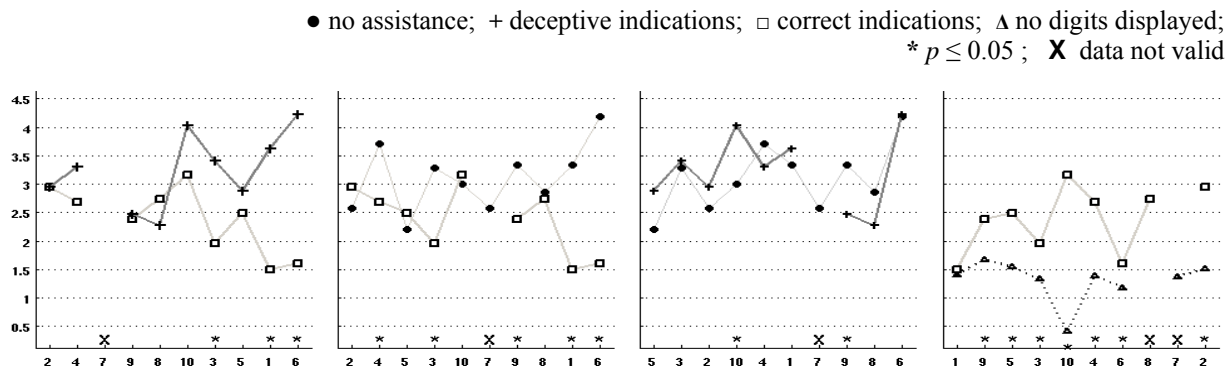


Figure 36: Comparing the number of cards inspected during the search for the correct target card. Conditions compared and order of subjects are the same as in Figure 35

Table 2: Information about age, sex and dominant eye (left or right) of subjects in order of their participation, as well as F and p -values as results of the ANOVA comparing the distributions of number of inspected cards pair-wise between conditions, subject by subject. Results that indicate significance are printed in bold letters.

Subject	age	sex	dom. eye	correct ind. vs. deceptive ind.		correct ind. vs. no assistance		no assistance vs. deceptive ind.		correct ind. vs. no digits displayed	
				F	p	F	p	F	p	F	p
1	23	f	r	36.3	< 0.01	38.8	< 0.01	0.2	0.622	0.1	0.731
2	25	m	r	0.0	0.942	1.1	0.307	0.7	0.404	27.0	< 0.01
3	24	f	r	16.9	< 0.01	14.8	< 0.01	0.0	0.873	7.7	0.008
4	25	m	r	2.7	0.108	5.7	0.021	0.4	0.513	20.6	< 0.01
5	24	m	r	0.4	0.529	0.5	0.485	1.9	0.176	8.2	0.008
6	25	m	r	46.5	< 0.01	39.7	< 0.01	0.0	0.830	4.1	0.050
7	26	m	r	-	-	-	-	-	-	-	-
8	30	f	l	1.1	0.313	0.0	0.972	0.7	0.410	-	-
9	33	f	r	0.1	0.777	8.7	0.005	6.9	0.012	5.3	0.026
10	25	m	r	3.9	0.055	0.2	0.657	5.1	0.030	114.1	< 0.01

Table 3: Results of questionnaire: Subjects had to answer questions with a rating on a 5 point MOS scale. Furthermore they were asked to choose which they considered as the fastest and as their preferred condition.

topic of question	mean rating on 5 point MOS scale			
realistic movements	3.3			
conscious of following correct indications	3			
conscious of following false indications	1.5			
precision of indications without digits	3.6			
speed with correct indications	4			
speed without assistance	3.9			
	choice of 10 subjects between four options			
	false ind.	correct ind.	no assistance	no digits on cards
fastest condition	0	4	5	1
preferred Condition	0	4	5	1

Summarizing the results of the first experimental series of the card game, it can be stated that there is a positive effect of the assistance given by the clone, but not for all subjects contrary to what was expected. The expectation that *deceptive indications* by the talking head would have a strong negative impact on user performance could not be confirmed. The results indicate that in this case subjects are able to ignore the talking head in a way that it does not irritate them. Their performance is then comparable to the condition *no assistance*. However, compared to *correct indications*, *deceptive indications* result in significantly longer reaction time more often than when no assistance is given. As the quality of the talking head is not rated very good, and some subjects uttered spontaneously disapproval of the repetitive utterance of encouragement given by the talking head, this was considered as an aspect of the setup needing improvement. This is also the case for the order of conditions, which was kept the same for all subjects and may have influenced the performance. Learning effects that might for instance have resulted in decreasing reactions times for succeeding cycles of a same condition or over conditions, could not be observed. The examination of gazes towards the talking head itself did not allow for any conclusions on performance or strategy. As its size on the screen is large, its gestures can probably be interpreted by peripheral vision. Explicit gazes towards the talking head can therefore not been used as a measure of the amount or degree of attention dedicated to it.

2.3 EXPERIMENT II: IMPACT OF MULTIMODAL DEICTIC GESTURES

The utterances of encouragement given by the talking head during the first series of experiments seemed to irritate some of the subjects. In fact utterances appropriate for the task should be short and clear, in accordance with the instruction to react fast given to the subjects. Furthermore they should contribute to attract the attention to the object of interest. These considerations have been taken into account for the design of experiment II. It aims at evaluating the possible benefit from the enhancement of the multimodal deictic gestures of the talking head by a spoken instruction. The deictic gesture is enhanced by a concomitant utterance as is commonly the case in real interaction.

The block-wise presentation of conditions was maintained. To compensate for eventual effects of order of presentation, the conditions were this time presented in random order. Fourteen users (ten male and four female) took part in this experiment. They ranged from 21 to 48 years of age and most were students. All regularly use a computer mouse and none reported vision problems. The dominant eye is the right eye for 8 subjects and the left eye for the other 6 subjects. When controlling the data for extreme outliers and validity, 11 of 336 measured reaction times had to be corrected. This corresponds to 3.3% of the measured values. Never more than two values had to be corrected per experimental condition of a subject.

2.3.1 Experimental conditions

The second series of experiments is based on experiment I and consists likewise of four different conditions. As a major difference the head and gaze movements of the clone are accompanied by the utterance of the demonstrative adverb “là!” (engl.: “there!”). This is an enhancement of the multimodality of the deictic gestures given by the talking head with audio-visual gesture. No equivalent utterance had been presented in experiment I where only utterances of motivation and felicitation were given between turns. These are not generated during experiment II.

To be able to compare the two experimental series, the condition *no assistance* was replicated as a reference. In the conditions *deceptive indications* and *correct indications*, speech onset is initiated 100ms after the onset of the deictic gestures. With the choice of this delay we tried to

approximate the temporal delay between eye movement and corresponding speech, measured by Bérar (2003). However, no exact information about the timing can be given. This is due to the imprecise control of timing for the triggering of events of the finite-state automaton used for the realization of the experiment and the not controllable scheduling of processes in the Windows operating system. The condition of the first series of experiments, where no digits were displayed on the cards, is replaced by the condition *correct indications with delay*. In this case an additional delay of 200 ms was introduced between the facial and the following acoustic deictic gestures.

2.3.2 Results

Before statistical evaluation, again the reliability of the measured data was examined as described above. During experiment II only one click error between neighbouring cards occurred (subject six in the condition *deceptive indications*). As can be seen in Figure 37, the analysis of the reaction time evidences an advantage for 7 out of 14 subjects of a mean of 340ms in the condition *correct indications* against *deceptive indications*. Comparing *correct indications* to *no assistance* this is the case for 8 out of 14 subjects with a mean gain of 360ms per cycle. This has to be put in relation with the mean reaction time over all subjects and condition of about 1.8 seconds. The proportion of users benefiting from this advantage is more important than they were in experiment I. Similar to the findings in experiment I, when comparing *deceptive indications* to *no assistance*, there are conflicting tendencies. Three subjects show shorter reaction times while three other subjects show longer reaction times when no assistance is given. A similar effect occurs in the comparison of the two different temporal alignments of the onset of vocal and visual gesture. In the case of the extended delay between the two onsets, two subjects show shorter reaction times while three subjects show longer reaction times.

For the analysis of the data collected with the eye tracker, subject 2 was completely excluded from evaluation due to insufficient monitoring of eye gaze, subjects 1 and 9 only partly. In Figure 38 this is marked with 'X' above the subject number on the x-coordinate. The over all mean of number of inspected cards is about 2.8 cards.

Analysing the data with an ANOVA for significance at $p = 0.05$, it was found that eleven out of the thirteen subjects with valid data looked on average at 1.3 cards less in the condition *correct indications* compared to *deceptive indications*. Comparing to the condition *no assistance* where no talking head is displayed the mean is 1.4 cards for ten out of the twelve subjects with valid data. This time for all of the subjects with sufficiently valid gaze data no statistical difference between the numbers of inspected cards could be found between the conditions *deceptive indications* and *no assistance*.

The additional delay between the onset of vocal and visual gesture produces no clear tendency. For two subjects it seems to be advantageous, for three others it means impairment. Answering the questionnaire, eleven of the fourteen subjects estimated that they have the best reaction times when correct indications are given by the talking head. Most of the subjects declare that they glance a lot at the talking head when it gives correct indications and discard gestures when not. Still they estimate that this would have only little influence on their reaction time. Again, the individual subjective estimation of performance does not systematically agree with the objective evaluation. The movements of the talking head are judged slightly more realistic than in experiment I. The topics treated in the questionnaire and the mean rating by the subjects on a five point MOS scale are given in Table 6 (see section 5.1 for exact wording of questions).

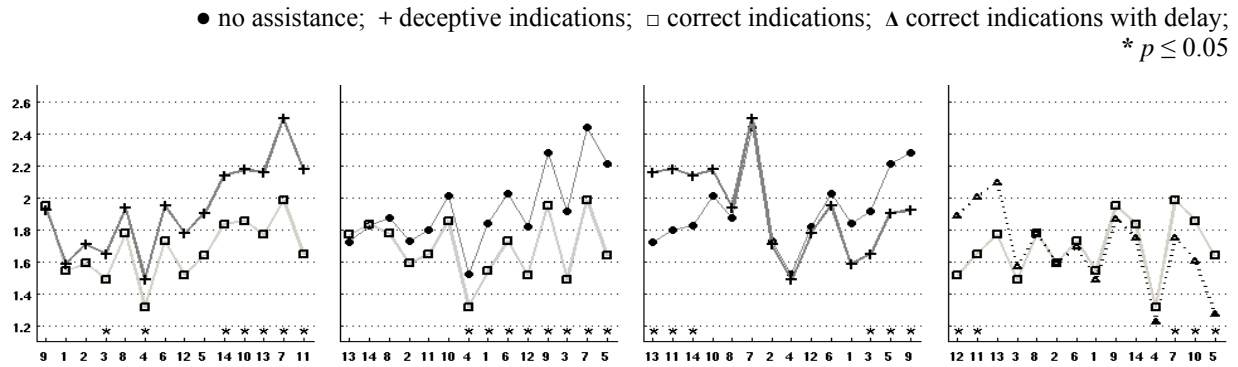


Figure 37: Comparing reaction times for four pairs of conditions. From left to right: condition 2 vs. condition 3; condition 1 vs. condition 3; condition 1 vs. condition 2; condition 4 vs. condition 3. The x-coordinate lists the subjects whereas the digit represents the order of participation. Mean reaction times for each user and for each session are displayed together with the statistical significance of the underlying distributions (stars displayed at the bottom when $p < 0.05$).

Table 4: Information about age, sex and dominant eye (left or right) of subjects in order of their participation, as well as F and p- values as results of the ANOVA comparing the distributions of reaction time pair-wise between conditions. Results that indicate significance are printed in bold letters.

Subject	age	sex	dom. eye	correct ind. vs. deceptive ind.		correct ind. vs. no assistance		no assistance vs. deceptive ind.		correct ind. vs. correct ind. delayed	
				F	p	F	p	F	p	F	p
1	22	m	l	0.1	0.764	5.0	0.031	3.0	0.088	0.9	0.356
2	22	m	l	1.5	0.227	1.0	0.313	0.0	0.987	0.0	0.893
3	21	m	r	6.0	0.018	15.3	< 0.01	4.4	0.041	1.6	0.207
4	30	m	r	4.6	0.037	12.1	0.001	0.3	0.583	3.4	0.072
5	21	m	r	3.9	0.054	19.4	< 0.01	4.4	0.042	16.5	< 0.01
6	48	m	l	2.8	0.103	6.5	0.015	0.3	0.587	0.3	0.585
7	27	m	l	15.1	< 0.01	10.7	0.002	0.1	0.709	11.5	0.001
8	37	m	r	2.5	0.122	1.2	0.280	0.3	0.569	0.0	0.877
9	26	f	r	0.1	0.807	4.2	0.047	5.0	0.031	0.3	0.596
10	26	f	l	9.2	0.004	2.2	0.144	2.6	0.112	6.3	0.016
11	25	f	r	17.8	< 0.01	1.6	0.216	6.5	0.014	13.7	0.001
12	28	f	r	3.3	0.078	6.5	0.014	0.2	0.658	6.1	0.017
13	29	m	r	5.8	0.020	0.2	0.696	7.3	0.010	1.7	0.194
14	29	m	l	4.1	0.048	0.0	0.925	4.9	0.032	0.4	0.549

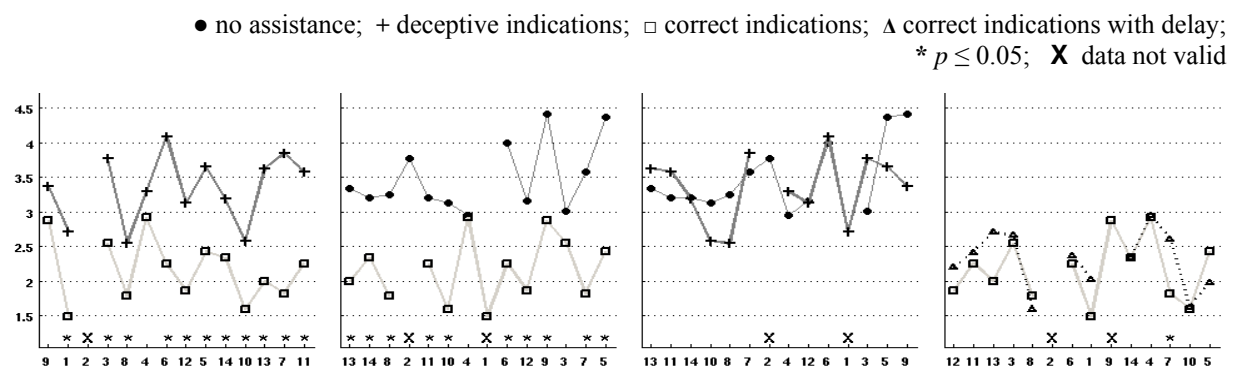


Figure 38: Comparing the number of cards inspected during the search for the correct target card. Conditions compared and order of subjects are the same as in Figure 37

Table 5: Information about age, sex and dominant eye (left or right) of subjects in order of their participation, as well as F and p -values as results of the ANOVA comparing pair-wise between conditions the distributions of number of inspected cards. Results that indicate significance are printed in bold letters.

Subject	age	sex	dom. eye	correct ind. vs. deceptive ind.		correct ind. vs. no assistance		no assistance vs. deceptive ind.		correct ind. vs. correct ind. delayed	
				F	p	F	p	F	p	F	p
1	22	m	l	12.2	0.001	-	-	-	-	3.7	0.060
2	22	m	l	-	-	-	-	-	-	-	-
3	21	m	r	11.5	0.001	0.5	0.476	3.6	0.063	0.0	0.887
4	30	m	r	1.0	0.324	0.0	0.831	0.7	0.410	0.0	0.874
5	21	m	r	6.9	0.012	21.9	< 0.01	2.1	0.151	2.4	0.126
6	48	m	l	20.8	< 0.01	18.1	< 0.01	0.0	0.826	0.2	0.664
7	27	m	l	20.9	0.000	15.0	< 0.01	0.4	0.527	11.6	0.002
8	37	m	r	4.2	0.047	13.5	0.001	2.2	0.145	1.1	0.310
9	26	f	r	1.9	0.176	8.3	0.006	2.3	0.134	-	-
10	26	f	l	8.0	0.007	16.7	0.000	1.3	0.267	0.0	0.906
11	25	f	r	15.6	< 0.01	5.2	0.028	1.0	0.318	0.1	0.712
12	28	f	r	8.6	0.005	11.9	0.001	0.1	0.794	1.2	0.271
13	29	m	r	13.0	0.001	6.3	0.016	0.4	0.507	3.2	0.081
14	29	m	l	7.0	0.012	4.4	0.042	0.1	0.758	0.0	0.896

Table 6: Results of questionnaire: Subjects had to answer questions with a rating on a 5 point MOS scale. Furthermore, they were asked to choose which they considered as the fastest and as their preferred condition.

topic of question	rating on 5 point MOS scale		
realistic movements	3.9		
conscious of following correct indications	4.2		
conscious of following false indications	2.2		
false indications are disturbing	2.6		
speed with correct indications	3.3		
topic of question	rating on 5 point MOS scale		
	choice of 14 subjects between three options		
	false ind.	correct ind.	no assistance
fastest condition	1	11	2
preferred Condition	0	11	3

2.4 DISCUSSION AND PERSPECTIVES

Summarizing the results from both experiments, it can be stated that deictic gestures used as positive assistance by the talking head are able to reduce processing time of subjects in a search and retrieval task. Furthermore, they are able to reduce the number of cards checked during the search process, which we interpret as a reduction of cognitive load. There may be more efficient modalities available than deictic facial gestures to reference objects in the given scene. This however was not the objective of the described experiments. The aim was rather to test the efficiency of the gestures once a talking head is present, which our experiments proofed successfully. However, the fact that subjects manage to ignore deceptive indications, suggests that the appearance of the talking head is not as dominant as it was expected to be for the directing of attention. In the condition in which deceptive indications

are given, a decline of performance was expected relative to the condition in which there is no assistance, but was not confirmed in the experiment. No major differences are observed between these conditions.

The detailed inspection of the data did not allow for the determination of general relations between the respective measured values. The measurements of processing time and number of inspected cards do not necessarily show the same tendency and especially the declarations made in the questionnaire are often not coherent with these objective measurements. The subjects seem to have used very different individual strategies. This is not astonishing as compared to experiments following the Posner paradigm, where the demanded reaction of subjects is typically a mere button click, our experiments needed a rather complex reaction, necessitating the movement of the mouse and its use for the selection of cards. Some subjects indicated that personal motivation may also have had an influence, such as the aim to outperform the talking head and not take its indications into account.

From the first series of experiments, considering the condition where no digits are displayed on the cards, we conclude that control and rendering of deictic gestures implemented in the talking head are sufficient to localize objects in space. For the choice between close neighboring objects, however, additional information is necessary to allow for a reliable decision. Without such additional information a sufficiently precise interpretation of the facial gestures of the talking head seem not to be possible. These findings gave place to a Master's project to model eyelid movements in dependence of gaze direction relative to head orientation (Casari (2006)). The model produces very good visual results and explains a large percentage of the deformations observed in the corpus (see Figure 39). It was however terminated only after the two experiments described here had been completed. It could therefore not be included for testing its impact on the precession of interpretation of purely visual gestures.

An important result is that a more refined manner of giving deictic cues as is promoted in experiment II is able to strongly increase the impact of the deictic gestures. The processing time as well as the cognitive load could be reduced. This is also reflected in the subjective judgement of movements of the talking head that are judged more realistic than was the case in experiment I. After the experiment II, there were no remarks of subjects to complain about the vocal instructions of the talking head as was the case in experiment I.

The objective measurement shows that a higher percentage of subjects profits significantly from the assistance given by the talking head. The gain in reaction time however can not be claimed to be different between the two series of experiments. Comparing the conditions *correct indications* to *no assistance* for reaction time the percentage of subjects that produce significant differences nearly doubled from 30% to 57%. In the case of number of cards inspected during the search it augmented from 56% to 83%. When comparing *deceptive* to *correct indications*, the percentage of subjects that show a significant difference between the two is 50% in both cases. Concerning the number of inspected cards it augmented from 33% to 85%.

The test of two different alignments of vocal and visual gesture in the second experimental series shows that this may have an influence on performance. Therefore a successful implementation of audio-visual deictic gestures in an application claims for detailed investigation of real human behaviour in order to choose an optimal temporal alignment. In such a context, also the exact coordination of eye and head movement and their trajectories should be investigated in detail. The movements simulated in the experiment were only coarse approaches to the kinematics known from literature.

The experimental scenario presented here could probably be further improved by displaying more objects on the screen and varying their size and the size of the digits. According to Fitts's law, distance and size of object are known to influence performance in pointing and

selection (Surakka, Illi & Isokoski (2003)). This would require a closer examination of the objects and increase the number of objects to check in order to find the correct one without the assistance of the talking head. Therefore the benefit of the assistance should become more prominent. However, we consider the results with the current implementation as sufficient confirmation of our assumptions and encouraging motivation to study further possibilities to enhance the capabilities of the talking head as an embodied conversational agent.

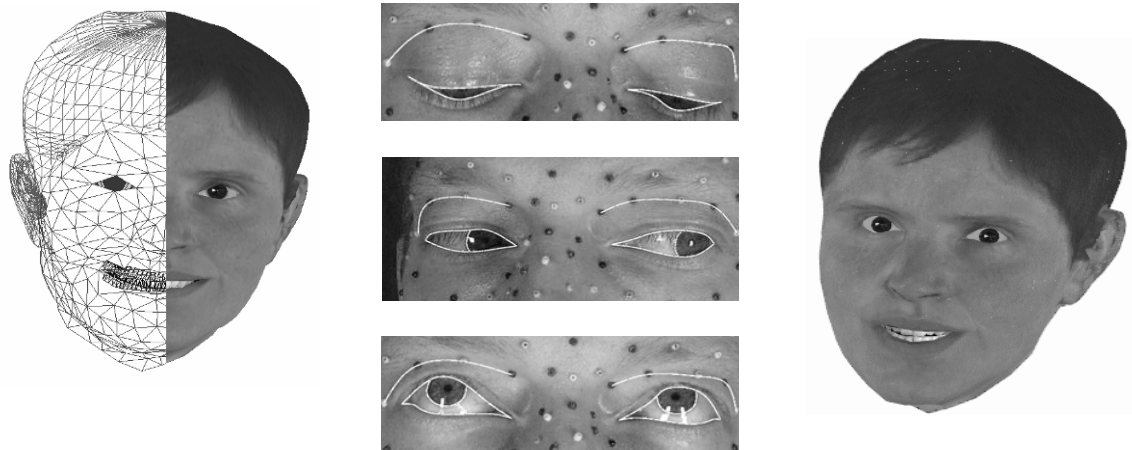


Figure 39: Evolution of eye model: Left: modeling of eyelids as two triangles described by four vertices to realize closing of eye and blinks. Middle: measurement of shape of eyelids in relation to the position of the pupil for the development of an advanced eyelid model. Right: implementation of eyelid model in the female talking head.

3 MEDIATED FACE-TO-FACE INTERACTION

The early work on gaze behavior in face-to-face interaction distinguishes between two conversational states, speaking and listening, and two directions of gaze as either directed towards an interlocutor or away (Kendon (1967), Argyle & Cook (1976)). Mutual gaze is mentioned as the special event when both persons in a dyad directly gaze to each other. In several cases, these categories are subdivided, as for instance into beginning or end of discourse or further categories are mentioned. There are however no clear definitions of sub-categories. Nowadays the improvements of technical knowledge provide researchers with the possibility to monitor - with high temporal and angular resolution - the direction of gaze as angles relative to the orientation of the head. The additional use of video cameras offers the possibility to take the scene into account and to associate gaze with distinct objects or targets. This claims for a resumption of investigations in this domain, with a finer grained classification of gaze direction as well as a more detailed structuring of interaction.

We studied dyadic mediated face-to-face conversations involving simple question-answer dialogues. The outcome should give insight into the relations between the different states that persons adopt in face-to-face interaction and the observed gaze patterns. We are interest in relations between states and gaze of each speaker, as well as mutual influence between the gaze behavior of two interacting subjects. Such knowledge is of great interest for the animation of embodied conversational agents (ECAs) to establish a better perception of presence of the ECA, to enable a grounded interaction and to signal mutual attention.

3.1 SCENARIO AND SETUP

The main interest in this experiment was the analysis of gaze behavior during face-to-face interaction. The experiments should unveil the influence of communicative activity of a person on its own gaze patterns. Furthermore it should clarify to what extent there is an influence of the interlocutor's behavior on gaze direction. Gaze behavior motivated by endogenous processes have a functional role for signaling and enlightening speech chunks, turn taking or enhancing other communicative gestures. Gaze behavior produced as reaction to the interlocutor's behavior results from the fast reactive loops of interaction that play an important role in the signaling of attention and the establishment of grounded interaction.

We developed a setup that provides the means to put to subjects into interaction in a way that enables natural or close to natural interaction and the acquisition of high-resolution data. The audiovisual appearance of the subjects as well as their gaze directions can be recorded in a way that enables the matching of stimulus and gaze direction in order to identify gaze targets. Furthermore, the setup provides temporal synchrony and consistency of the recorded audiovisual and gaze data.

The scenario used for this experiment favors a recurrent occurrence of events we consider as important in such dyadic interactions, in order to obtain a sufficient number of samples that can be used for statistical analysis.

3.1.1 Scenario

The experiment on face-to-face interaction described here is clearly inspired by the work of Kendon (1967), Argyle & Cook (1976) and others, dedicated to the analysis of gaze behavior in interaction and dialogue (see chapter 1.4, page 44). It is however the first experiment known to the author, to investigate gaze behavior with close to natural conditions and objective technical measurement at such a fine degree of resolution.

The commonly used segmentation of dialogue into intervals of speaking and listening has been further subdivided, notably distinguishing between onset and main parts. This should simplify the determination of relations between segments of conversation and gaze patterns. In previous studies, authors sometimes distinguished between end or beginning of speaking or listening, but without clearly defining the boundaries (see Kendon (1967)). Casell et al. (1999) furthermore distinguished between ‘theme’ and ‘rheme’ (see section 1.4.3.4).

In order to obtain sufficient and statistically significant data for each of the defined segments of interaction, we assure a sufficient number of events of the same type. At the same time, it is desirable to avoid the appearance of single events not able to be classified in the desired context. We imagined an experimental scenario that fulfils these constraints without inhibiting natural interaction. We chose the person that served as model for our female talking head as target subject that participates in all recorded interaction (female, French, researcher, dominant right eye, right hander). This allows to acquire a relevant amount of data for a single subject and to level out eventual peculiarities in her behavior that are due to the personality of her interlocutor.

As social norms, sex and relations of hierarchy are known to have major influence on gaze behavior (Argyle & Cook (1976)), we considered these factors for the definition of the scenario and especially the choice of subjects. The subjects we put into interaction are all of the same sex (female) with similar social background and status (researcher, French).

To enable a continuous monitoring of the subject’s gaze behavior, the scenario must impose restrictions on the subjects to limit their head movements to the space surveyed by the measurement equipment. Restricting devices such as headrests are not appropriate in this context, as they would too much interfere with a natural appearance of the subjects’ faces. We imposed the necessary restrictions with the design of the scenario and the task, without demanding a conscious effort to the subjects that might have undesired influence on the course of interaction.

3.1.1.1 *Definition of conversational role and cognitive state*

For the segmentation of the interaction, we considered that there are two basic roles a person can adopt in interaction, independent of who has the turn. One person may lead the interaction, deliver the topic and content and direct the course of the conversation. The interlocutor on the other hand takes then a more passive role, tending to follow the interaction and to respond to questions and proposals. We tried to take these differences into account in the experiment as two different roles that we defined as ‘*initiator*’ and ‘*respondent*’. The *initiator* is the person that dominates the interaction, gives new information, defines the topic and listens with the intention to verify that the other is following but not necessarily in order to receive new information. The *respondent* is receiving new information and is reacting to the other person. These roles can change recurrently during the interaction. Independent of role, numerous changes of turn may appear in between. The scenario of the experiment should allow a clear distinction of these roles.

For the further segmentation of the discourse, we defined ‘*cognitive states*’ (CS) as classification of the segments of the discourse. In association to ‘mental states’ as mentioned by Baron-Cohen (1995), *cognitive states* characterize the mental activity of the subject. We considered perceptible motor activities causing changes or maintenance of these cognitive states. On the one hand, this is necessary to provide the experimenter with clues to segment the dialogue. On the other hand, this is important in the context of mutual influence on gaze behavior. The segments describing the *cognitive states* of one person should be appropriate to establish a relation between these states and eye gaze events observed either sides of the interaction.

3.1.1.2 *Distinction of cognitive states*

For the segmentation of our experimental data we distinguish the cognitive states ‘*speaking*’, ‘*listening*’, ‘*waiting*’, ‘*thinking*’, ‘*reading*’ and ‘*pre-phonation*’. Intervals that cannot be attributed to one of these states or that are not of interest for the analysis are classified with the further category ‘*else*’. Similar to the definition of cognitive states, this choice is also based on reports from literature.

Garau *et al.* (2001) distinguished the two modes ‘*while speaking*’ and ‘*while listening*’ for the animation of the gaze of avatars. This distinction again is based on the research reported by Kendon and Argyle (see section 1.4.3, page 47). As no further modes such as waiting or thinking are reported by Garau *et al.* (2001), the states they use correspond most probably to a wider scope than the *cognitive states* of *speaking* and *listening* used here. These are closely associated to the presence of a speech signal by one or the other subject.

When subjects are idle, waiting for the other to speak, we label this as ‘*waiting*’. Usually it occurs while the interlocutor is ‘*reading*’. According to the task, subjects had to read sentences from a paper in order to utter them in the following, addressing the interlocutor.

Minato *et al.* (2005) monitored the gaze behavior while subjects were thinking about the answer to a question, which corresponds closely to the definition of the CS *thinking* used in our own experimental work. They defined it as the interval between the end of a question and the beginning of the answer. In our experiments ‘*thinking*’ is the interval during which the subjects prepare a sentence in mind.

When subjects articulate, but do not yet produce audible speech we label this interval as ‘*pre-phonation*’. Usually it contains clearly visible mouthing, produced to prepare for speaking.

The recorded interactions are segmented into these categories without leaving gaps between the respective intervals. Inside the analyzed interactions, which are delimited by the respective beginning and end of the intervals of *conversational role*, only few intervals are labeled as ‘*else*’.

3.1.1.3 *Faked interaction using a pre-recorded stimulus*

To have the possibility to unveil the consequences of action-perception loops during the interaction, the scenario should enable the use of prerecorded stimuli. It should allow the replacement of the audiovisual stimulus of our target subject by the replay of a recording. The data acquired during such faked interaction can then be compared to the data acquired during direct online interaction. This should clarify the effect of the absence of a reactive loop while preserving all other aspects of the stimulus, such as naturalness and credibility.

To generate the stimulus for the faked interaction, an online interaction of our target subject with one of the experimenters was recorded on VHS video tape. As this stimulus is taken from a real interaction, the gaze behavior of our target subject, her facial gestures, her rhythm of utterances and the pauses in between where her interlocutor has to repeat the given utterance are entirely realistic. To explain the inconsistencies that may appear during the interaction, the subjects were informed that the aim of the scenario was to study the influence of visual feedback on mediated face-to-face interaction, and that the audio feedback was interrupted on purpose. They were not informed that they interacted with a recording. The fact that some of the subjects did not realize that they were watching a pre-recorded video, shows that the scenario meets our expectations.

3.1.1.4 *Control of interaction and dialogue to produce recurrent events*

A free discussion of a given or arbitrary topic would not have been appropriate for several reasons. It may engage the subjects too much in a personal or emotional discussion that may provoke ample body movements and generate complicated emotional states that were not planned to be the object of this study. The engagement of the subjects would probably lead to

struggle for the turn, with overlapping speech and multiple interruptions of the discourse. It may as well fail to produce a sufficient amount of data if the subjects do not manage to start or maintain a lively discussion.

Sentence building game

A first attempt was made with a sentence generating and repeating game. Therefore one person has to exactly repeat the sentence previously uttered by the other and then to continue the sentence with a short extension. In the following this extended sentence has to be taken over again by the first person, be repeated and extended anew. This results in increasingly long and complicated sentences, until one of the persons makes an error, not correctly repeating the given sentence. As this generates irregular durations of utterances, laughter and pauses in between, this proved not to be appropriate for the use of prerecorded stimuli.

Sentence repeating game

We modified the sentence building game into a sentence repeating game with a list of given sentences. The sentences are written on a sheet of paper. The speaker reads, memorizes and then recites each sentence to the interlocutor, one after the other. The interlocutor is asked to memorize and repeat them as exactly as possible in a single attempt. As motivation, the subjects were informed that the score of successfully repeated sentences would be determined. This scenario proved to have several advantages. The fact that the sentences are to be read from a paper that lies on the table in front of the monitor, prevents the subjects from changing too much the position relative to the screen or to retreat progressively. The defined beginning and ending of the sentences avoids uncontrolled turn taking. We expected that the cues signaling a change of turn would still be apparent, but that there will be neither conflicts nor overlaps of turns. Furthermore, such a scenario generates a very repetitive course of the interaction. This is advantageous to produce a controllable amount of classifiable states during the interaction, and favors automatic labeling.

The restrictions such a scenario imposes limit of course also the naturalness of interaction. The annotated cognitive states however are observed in the context of human interaction and do carry the basic characteristic of these states which is exactly what we are interested in. Only if the essential characteristics of these states are understood, the influence of further subcategories can be isolated.

Linguistic content

As we were especially interested in gaze behavior that signals interest and attention, we chose semantically unpredictable sentences – ‘*SUS*’ (Benoît, Grice & Hazan (1996)). These are grammatically correct sentences, but without any reasonable meaning. Therefore, the individual words cannot be reconstructed using top-down semantic information. This forces the subjects to be particularly attentive while listening to the interlocutor.

3.1.2 Setup for mediated face-to-face interaction

The setup should put two subjects into face-to-face interaction with the possibility to monitor gaze direction and audio-visual appearance without disturbing the interaction. Especially the fast reactive loops of interaction should not be impaired by the setup. We use eye trackers for the recording of gaze direction and video cameras and microphones for the audio and video recording. Figure 40 shows a schematic diagram of our setup.

Each subject is monitored with a video camera, a microphone and an eye tracker. The signals are recorded, and audio and video signals are sent to loudspeakers and a screen, where they serve as stimuli for the other subject. A very beneficial aspect of this setup is the possibility to

use prerecorded audio and video signals that can be presented to the subject as stimuli. Figure 41 shows the modified setup as used for experiments where prerecorded audiovisual stimuli are used. A criterion for the recorded data streams, important for the later analysis, is that they can be put in exact relation to each other, in order to associate events of different modalities. We explain in this section our choice of material, the experimental setup and the different procedures that have to be performed in addition to the experimental recording in order to enable valid data. The schematic diagram in Figure 42 describes the succession of the different steps of an experimental recording. Once the setup is adjusted to the subjects and eye trackers are calibrated, synchronization procedures are started for both subjects remotely from a third machine. Then reference targets are presented to the subjects by the means of a laser pointer. This is used to check the correctness functioning and calibration of the eye tracker. Then follows the actual experiment. At the end of the experiment, the control and synchronization procedures are repeated. This done to detect drift during the recording.

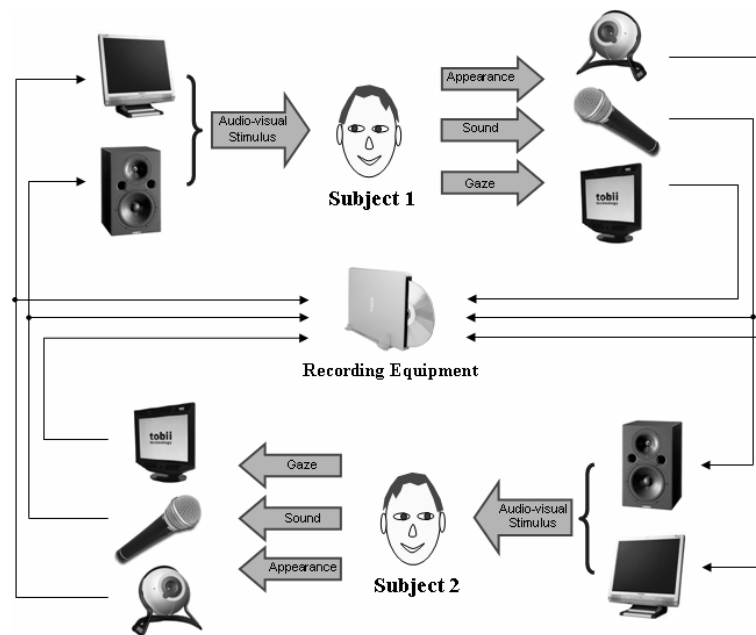


Figure 40: Setup for mediated face-to-face interaction of two subjects. The audio-visual appearance of a subject is captured with a microphone and a camera and presented to the other subject with loudspeakers and a computer screen. The signals are recorded along with the gaze data acquired with the eye trackers that are integrated into the screens.

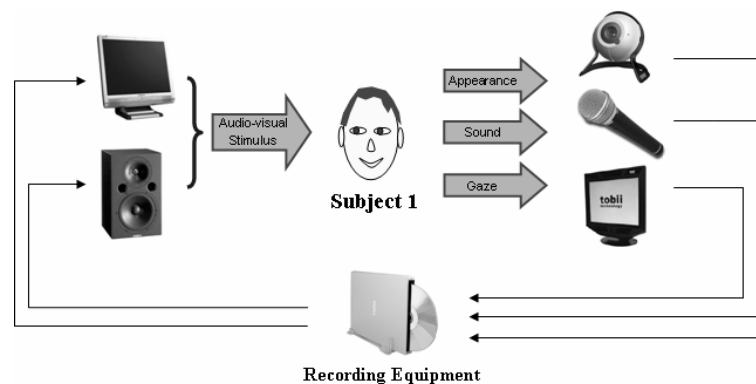


Figure 41: Setup for mediated interaction of a subject with a prerecorded audiovisual stimulus. The stimulus is played back from a VHS audio and video recording. The played back audio signal is recorded again on a separate channel along with the audio signal of the present subject, which enables temporal alignment of the data after the experiment.

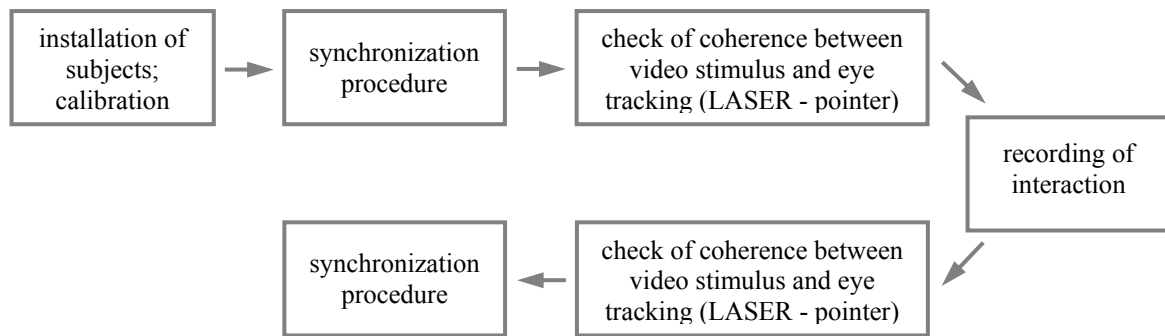


Figure 42: Chart of different procedures performed in connection with an experimental recording. The check using the LASER-pointer is not performed when using a pre-recorded stimulus during the faked interaction.

3.1.2.1 Choice of eye tracking system

For the eye tracking there were two options: a head mounted or a desktop system. A head mounted device has the advantage of free space of action. The measured gaze direction can be related to the scene by the use of a head mounted camera for video recording. The calibration of such a system is however very complicated as the angle of view of the camera is different from that of the eyes. A disadvantage of a head mounted system is that it occludes parts of the face of the person wearing it and that it would probably attract the attention of the interlocutor. Furthermore, it does not offer the possibility to replace one of the subjects with an animated agent or a faked stimulus, such as a video recording. For these reasons, a head mounted system is no practical solution for our purpose, although it allows for close to natural face-to-face interaction with direct contact in a common space of interaction. We chose a Tobii® 1750 system that integrates the eye tracker into a computer screen.

3.1.2.2 Choice of technique for video linkage

To assure an appropriate quality of data acquisition and presentation of video signal at the same time, we considered two techniques. We planned either to build a teleprompter system or to place a very small camera in the line of sight between the screen and the subject.

To generate a direct front view of the face, the camera should be straight in front of it, in the line of sight. A displacement of the camera away from this center position would generate the impression of a view from an angle onto the subject, even when facing and gazing at the screen directly. Considering displacements of the camera away from the ideal position, people seem to be less sensitive when the camera is slightly moved further upwards (Chen (2002)).

A possible solution is the use of a half-silver mirror as typically used in teleprompter systems. Mukawa *et al.* (2005) used such a setup that they describe as ‘video-mediated communication system’. Garau *et al.* (2001) used a similar setup that they called ‘video tunnel’. A teleprompter system uses a half silvered mirror, which lets pass a certain percentage of light and reflects the rest. Behind the mirror that is inclined by 45° relative to the line of sight, a camera is placed to capture the passing light and to film the scene. If a dark background is behind the mirror in viewing direction, images can be projected onto the mirror from above or below, which are reflected into the direction of the observer. The observer will perceive them as if they appeared straight ahead on the background. This setup enables the use of any camera, and a high quality camera of big size could be chosen.

For several reasons we did not choose this option. Most important is the restriction imposed by the eye tracker. As we intended to measure the gaze direction of the subject, we used Tobii® eye trackers that are integrated in the outer frame of the casing of a computer screen. A mirror placed between the screen and the subject would have hindered the functioning of

the eye tracker. Furthermore, an in-house realization of a teleprompter system would be a complicated mechanical and very time consuming work. To exploit the possibilities offered by such a setup, the purchase of expensive high quality cameras would have been necessary. The purchase of two teleprompters would have been too expensive. Such costs would have been exaggerated for the purpose of a first precursor experiment, addressing the given topic without the guaranty of valuable outcomes.

For these reasons, we decided to use a setup where the camera is placed directly between the observer and the screen. In this case, however, the camera covers a part of the screen. In order not to disturb the view, the camera needs to be well positioned and to be of minimal size.

3.1.2.3 Positioning of camera, angle of view and distance

Compared to a direct live interaction, the mediated interaction via the crossed screen-camera setup implies several peculiarities that may influence the interaction. First of all the three dimensional appearance of the subjects is reduced to a two dimensional display on the computer screen. When one subject moves relative to the screen, the view of the displayed subject will not change. In real direct interaction, this would result in a more profile view of the other person. Stereoscopic displays in combination with stereo video acquisition propose a future solution for this problem. Head mounted stereo displays and stereo cameras (e.g. stereo vision camera Bumblebee²) already exist.

Another problem is to generate a view that produces an appropriate perception of distance. When looking at a photograph, we can estimate the difference between the person taking the picture and the person being photographed. This is possible even independent of the size of the photograph, the proportion between the size of the head and the outer frame and also independent of the ground being visible or not. We hypothesize that this information can be taken from the relation between front and side view of the head, and other related deformations of the view. A close view for instance hides the sides of the head such as the ears, whereas a distant view brings them further into view.

As both subjects are seated at the same distance from the screen of about 50cm, the video presentation should generate the impression of a distance of about 1m between the two subjects. In the final arrangement, we placed the camera directly on the screen and adjusted the position of the subjects so that their image on the other screen will make the camera appear slightly above the bridge of the nose. This is estimated to be the best position to enable straight ahead perception of the other subject. At the same time, we consider this as the best distance for the employed cameras, to generate an appropriate impression of distance. We use mini-cameras of cubic with a side length of about 2cm. This allows their positioning on the screen without occluding important parts of the displayed scene.

We must admit that problems may arise concerning the perception of saccade angles that might not be reliable. The studies on the perception of gaze direction however justify the assumption of sufficient reliability for our purposes (see section 1.3.2.1, page 32). The problems of the two dimensional appearance are inevitable, but are also imminent to the use of an embodied virtual agent that is limited to the same restrictions when visualized on a computer screen.

To eliminate major problems from the described arrangement of the setup, we verified the perception of gaze under these conditions. We engaged two persons, who saw each other via the experimental setup, in a short test. Colored cards were inserted into the video image in order to add further objects to the scene (see Figure 43: scene as perceived by subjects). While one subject fixated a target on the screen, the other had to estimate the fixated target as either a virtual target on the screen or a real target on his own face. Both the looker and the observer

² <http://www.ptgrey.com/products/stereo.asp>

marked the actual and estimated gaze targets on a schematic drawing of the scene (see Figure 43, left). The role of ‘looker’ and ‘perceiver’ was further exchanged. The results show that when using the experimental setup coherent perception of fixated targets is possible. This encouraged us in our assumptions on the establishment of a veridical representation of subjects and a setup allowing for reliable measurements.

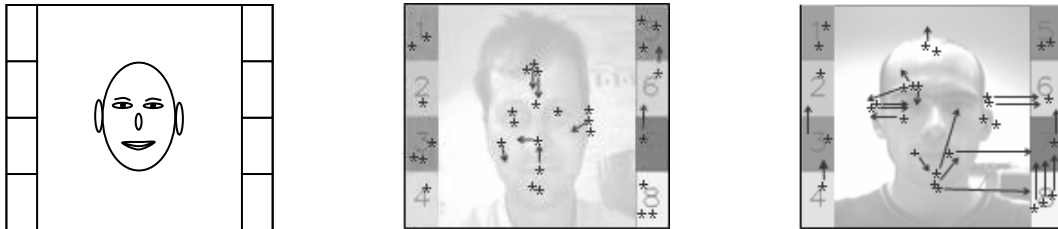


Figure 43: Comparison of perceived and actual target of fixation. Left: Schematic drawing of the scene for the marking of actual or estimated gaze target. Middle and right: The stars mark the fixated targets of the observer. Arrows indicate the location that the observer (person shown in picture) estimated to be the fixation target, if there is a discrepancy between his estimation and the actually fixated target.

3.1.2.4 Recording device (VHS, dvd, hard disc)

Along with the Tobii® 1750 eye tracking hardware, Tobii® proposes software for the measurement and evaluation of data. ‘ClearView®’ is a graphical interface for the configuration and control of the eye tracker. It provides users with a calibration procedure, the choice between different modes of stimulus presentation as well as the recording of external signals and various tools for the post-processing of data. ‘TETserver’ is a software tool for direct online access to the measured gaze data. It allows for the integration of the gaze data acquisition in software applications.

The ClearView software proposes the possibility to record external video signals in synchrony with the gaze measurement. The videos can be exported with synchronously overlaid information about gaze targets. As no information about the reliability of synchrony of data was available, this was tested with a laborious procedure.

Audio-Video recording with ClearView

To test the temporal reliability of recordings made with ClearView we needed a reference for comparison. For this purpose, we used a Panasonic NV-HS900 VHS video recorder. As it is an analog device, there is no buffering of data. Digital device in contrast need buffering for the coding and decoding of data that may lead to suppression or delay of data. Temporal distortion or delay in VHS recording can only originate from unstable speed of the actuation during recording or playback. As there was no reason to presume such instability, we considered the VHS recording as a reliable time basis.

Our laboratory has of a reliable acquisition software (DPS Velocity V8) that we used to digitalize the VHS recordings. The video exported with ClearView could then be compared to the VHS recording for frame duration between events.

When comparing the two recordings we found differences of several frames between events that made the video exported from ClearView shorter. As the raw video recording is accessible in the file system of the software, we compared this to the exported version. It was possible to identify frames that existed in the raw video data, but were missing in the exported version. The problem was identified to arise from the process of super positioning of gaze data onto the video frames during the export procedure. We could not clarify, if the gaze data was subject to the same loss of data. This might have resulted in consistency between video frame and corresponding gaze target in spite of the loss of data. It could however not be excluded that a temporal mismatch of gaze data and video data was generated by the export

procedure. In any case, the videos generated with ClearView were not adequate for further analysis. Only the raw video recording accessible via the file system of the software proved to be reliable.

Audio-Video hard disc recording and recording software

Due to bandwidth constraints the available hardware, the recording of the second subject could not be made on the same computer to which the eye tracker was connected. We considered VHS video recording, a consumer DVD recorder and recording on a computer hard disk as alternatives. For the latter, VirtualDub³, which is an open source video capture and processing utility, was chosen as recording software. Experimental tests with the help of the VHS video recorder as described above demonstrated the reliability of the hard disk recording.

3.1.2.5 Synchronization of data

The synchrony between the different data streams is important for correct association of causal relations between events. As ClearView did not guaranty reliable synchrony between the different data streams, a synchronization procedure had to be developed especially for the purpose of our experiment.

Markers that are written to the different data streams during the synchronization procedure, are used to time align the different data streams for further treatment. We chose a LED light signal as a marker for the video frame and a beeper as marker for the audio stream. At the same instance, a marker has to be added to the gaze data recordings.

As the synchronization signals are triggered separately for the two subjects, the respective markers correspond to two independent reference time stamps. To be able to put them into relation, the audio signals of the two subjects are recorded as separate channels of the stereo audio recording. The temporal interval between the markers on the respective audio channels serves as a reliable reference for the temporal alignment of the corresponding data streams.

The recording provided by the ClearView software is not accessible during data acquisition. It is therefore not possible to add a marker to this data. TETserver in contrast provides the measured data online. A marker can easily be added to this data when it is written to a file. As the actual data of the experiment is acquired with ClearView, both ClearView and TETserver are run simultaneously during the synchronization procedure. This results in two data files containing the identical gaze data for the duration of the synchronization process, whereas the data recorded with TETserver in addition contains the synchronization marker.

As the two recordings contain identical data, they can be matched to determine to which time stamp in the ClearView file the synchronization marker from the TETserver file corresponds (see Figure 45). As ClearView runs continuously until the end of the experiment, the transposed marker is valid for the entire data stream.

To reduce temporal delay, we used the parallel port of the computer to trigger the synchronization signals instead of using sound and video signals that could be generated using the operating system. The parallel port is unbuffered and hence a more direct and reliable interface. An impulse is send to the parallel port of the computer at the same time as the marker is triggered to be written to the TETserver file. An active circuit, generating a short 5V rectangular impulse for a LED and a piezzo beeper when receiving the trigger impulse, is connected to the pins of the parallel port. The duration of the rectangular impulse is adjusted to 40 ms, which corresponds to the duration of a video frame. This enables the determination of the light burst onset in the video frames at a higher precision than frame rate.

³ <http://www.virtualdub.org/>

The synchronization procedure is started with a remote signal from a third computer, which triggers the start of the synchronization procedure for both subjects at the same time in order to reduce time shift between the respective synchronization markers. It starts the synchronization procedure, the Clearview recording of gaze data, TETserver and the video and audio recording.

During the synchronization procedure, the VGA source of the eye tracker screen is switched to the computer, which displays an image containing a yellow square. Once the square is fixated, it starts to execute a circular movement around the center of the screen (see Figure 44). When it crosses the vertical midline of the screen for the second time, the synchronization signal is triggered. The video, displayed on the screen, is important as a stimulus for the production of valid gaze data by the respective subject. This is needed for the matching of data recorded with ClearView and TETserver. To assure a correct recording of the video and audio synchronization signal, the beeper and LED is positioned at the level of the chin of the subject. Once the synchronization procedure is finished, the VGA source of the eye tracker screen is switched to the video signal from the camera filming the other subject. At the end of the recording, the synchronization procedure is repeated. This is used to compare the time span between the respective signals and to check the eventual drift during the experiment.



Figure 44: Stimulus presented on the screen during the synchronization procedure: left: initial image; right: indicated circular movement of gaze target to generate smooth a sequence of valid gaze data.

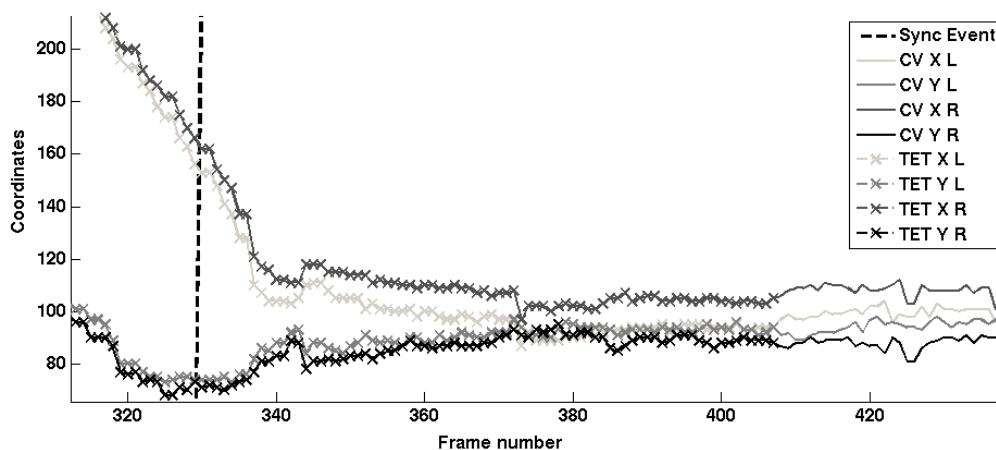


Figure 45: Example for the matching of gaze data. The X and Y coordinates for the left (L) and right eye (R) are displayed for the data streams recorded with the two different software tools. As expected, the respective curves can be matched and superposed. 'CV' stands for the data acquired with Clearview that is represented by continuous lines. 'TET' stands for the data acquired with TETserver represented by 'X'. The synchronization event is marked with a dashed vertical line.

3.1.2.6 Consistency between displayed and recorded video

We considered the possibility that the recorded video signal might not exactly be the same as the one displayed on the computer screen. Namely distortion or trimming of the borders might occur.

Filming a reference scene with markers, we verified the consistency of the video image from visual observation of the video, displayed on a screen. The markers were arranged to indicate the corners of the direct presentation of the recorded video signal. The video signal was recorded at the same time. In the playback of the video recording we controlled the consistency of the appearance of the markers in the corners of the video frame. Once the settings of the video acquisition card had been adjusted correctly no inconsistencies between the video images could be observed.

3.1.2.7 Consistency of video image and eye tracking

During the experiment, the video stimulus is not generated by the ClearView software. It is taken from the camera attached to the screen and a video console is used to transform it into a VGA signal for the display on a computer screen. As the video signal is not controlled by ClearView and originates from a different camera than the one used by the eye tracker, discrepancies between the display on the screen and the video information acquired and processed by the eye tracker might have occurred. In consequence, the coordinates of a gaze target as it appears on the screen might differ from the coordinates calculated by the eye tracker.

The video signal acquired with the internal infrared camera of the eye tracker is not directly accessible. We could therefore only verify visually if the gaze coordinates determined by the eye tracker matched the gaze targets of reference stimuli presented to a subject during a recording. For this recording, the cameras and screens were connected in the same way as during the experiment. A template with markers was presented at the distance at which the subjects would be placed relative to the camera. The markers served as visual targets. Four markers were placed in order to appear close to the corners of the screen. Another four markers were placed in the center, imitating the eyes, the nose and the mouth as possible gaze targets. This template was filmed with the camera and displayed on the eye tracker screen as video stimulus for a subject that was asked to fixate the targets, one after the other. The comparison of the gaze targets as measured with the eye tracker to the visual targets presented on the template was satisfactory.

In preliminary recordings of face-to-face interaction made to test the setup and scenario, we observed accumulations of fixations that clearly corresponded to the well-known triangular pathways between the eyes and the mouth (Vatikiotis-Bateson *et al.* (1998), Lansing & McConkie (1999)), but slightly shifted relative to the location of these targets in the video stimulus. To test if this was due to the precision of the eye tracker or its calibration, a procedure is conducted at the start and at the end of every experimental recording. This procedure reveals if mismatches between gaze and video data occur and whether they are constant over the duration of the experiment. An eventual systematic error in the form of a mismatch of data that appears at the beginning and at the end of the recording can then be corrected.

With a laser pointer three small light points are thus generated on the background screen behind one subject. This scene is displayed on the other subject's screen via the camera of the experimental setup. The light spots serve as locally restrained visual stimuli that are sufficient to trigger exogenous saccades, due to the light contrast, the color contrast (red on green) and the slight movement of the light point in consequence of the instability of the hand posture. In addition, subjects were informed that they should fixate the light spots. The setup was not altered between this procedure and the experiment itself. The respective frames are extracted from the video stream after the recording. No experimental data had to be discarded due to this procedure.

3.2 DATA PROCESSING

During the experiment, different devices and formats are used for the recording of data. Several procedures are developed to check for validity of data and to provide the parameters necessary for further treatment and analysis of the data, such as the synchronization procedure. In this section, we detail the procedures involved in the recording during an experiment and the manual treatment of the data necessary before it can be treated automatically with scripts developed for this purpose.

In this section, we explain in detail how the parameters for the further processing are extracted from the recorded data, and how the data is then processed automatically. The processing is mainly realized in Matlab scripts. A main script calls different Matlab functions, Perl scripts or the operating system to execute external programs. It is designed to separate the input of data values or graphical input from the actual processing of the data. Such data input are necessary for the interactive positioning of reference points for the tracking of the eyes and the mouth, or the demarcation of areas for the different regions of interest on the face.

3.2.1 Extraction of parameters from video signal

After the recording is finished the synchronization markers have to be extracted manually from the different files. Concerning the video recording, this is done with VirtualDub which allows frame by frame examination of the video stream. As the LED light signal lasts for 40ms, which is exactly the exposure time of a frame, the onset of the signal can be estimated more precisely than in the steps of the frame duration of 40ms. In certain cases, the signal is only visible on one frame. In others, it is distributed over two frames. From the differences in light intensity of the signal in these two frames, the onset is estimated to decimal fractions of frame duration.

Boundaries of the interesting parts of the interaction are determined from visual inspection of the video streams. For analysis of fixations, a reference image is chosen from one of the video frames. This frame should show a good representation of the subject's eyes as it will be used to initialize a software tool for the tracking of the eyes and the mouth on the video.

If a prerecorded stimulus from video tape is used, the described procedures are only necessary for the participating subject. At the beginning of the video playback, a synchronization sound signal is given, that is recorded along with the sound from the subject. This again is used to temporally align the experimental recording with the prerecorded data streams.

3.2.2 Extraction of parameters from audio signal

The onset of the synchronization signal in the audio stream is determined with the aid of Praat⁴. We visualize the intensity and spectrogram of the audio signal to determine the onset of the beeps. As the two audio signals from the two subjects are recorded on separate tracks of the same stereo recording, they share the same time scale. The time span between the onsets of the two synchronization signals is used to time-align the video and gaze data streams. From the intervals between the synchronization procedure at the beginning and those at the end, it can be determined if temporal fluctuations occurred between the different data streams. None of such inconsistencies occurred in any of the recordings.

3.2.3 Preprocessing of gaze data to generate data with constant frequency

The gaze data as recorded with ClearView includes the coordinates of the gaze target on the screen, a factor of reliability and a corresponding time stamp. It is supposed to be measured at

⁴ <http://www.fon.hum.uva.nl/praat/>

a constant frequency of 50 Hz (intervals of 20ms). The observed duration of intervals between data points usually varied between 19ms and 21ms. In rare cases, the intervals extended to over 100ms. The calculation of the mean duration of intervals from the given time stamps showed that the frequency of data acquisition does not correspond to the value of 50Hz indicated in the ClearView manual. We thus re-sampled the data at a constant frequency.

For the re-sampling, the moment in time from which the recorded data will be analyzed is set as the new reference zero point. From the new zero point on, the time stamp is increased in steps of 20ms until the end of the part to be analyzed is reached. The remaining gaze data before this time is not needed for later analysis and discarded. To every time stamp, the nearest neighbor from the original gaze data is assigned, if it is not more than half of the interval period (10ms) apart. In case that the superior and inferior time stamp are at the same temporal distance from the new time stamp, the data from the inferior time stamp is assigned. If no data can be assigned as no time stamp with valid data is in reach, the current time stamp is marked as containing invalid data.

The employed algorithm ensures that the coordinates of gaze direction are maintained as originally measured. Interpolation would have produced artifacts in the coordinates which may introduce serious errors as the jerky movement of saccades does not produce smooth curve when sampled at 50Hz. Interpolation between two consecutive data points for instance, that belong to fixations at distant locations, would result in coordinates indicating a location in between that is far from both fixations. The algorithm we employ solely affects data with an additional temporal imprecision of up to 10ms, whereas the coordinates as such are veridical. According to the information given in the user manual of the Tobii® eye tracker, the temporal accuracy of the time stamps in the measured data is ± 3 ms. After the re-sampling procedure accuracy must therefore be estimated to be ± 13 ms. We consider this as acceptable for our purposes.

3.2.4 Fixation detection

In human vision, visual information is mainly acquired during fixations and smooth pursuit movements. Saccades are executed to shift gaze rapidly between successive targets. We perform fixation detection on the raw gaze data. No obvious events were detected that may trigger smooth pursuit.

3.2.4.1 Fixation detection and visual processing

Fixations are periods of relative stability of gaze orientation that are of high importance in human visual perception. There is however no clear definition of the beginning and end of visual processing relative to the boundaries of a fixation, nor is there an absolute criterion for the determination of boundaries of a fixation. We are not able to relate fixations directly to definite cognitive processing. Furthermore, it must be kept in mind that the target of human attention is not identical with the gaze target (Langton *et al.* (2000), Simons & Chabris (1999)), but that both are closely related. It is however possible to detect periods of relative stability, that can be associated with fixations and the segmentation of gaze into fixations is a valuable step for the extraction of information from measured gaze direction. It informs about targets of interest and about the time that subjects dedicate to the inspection of these. An important aspect of fixation detection is furthermore to exclude gaze samples that are part of saccades. During saccades, visual perception is assumed to be strongly restricted. Gaze points that are measured during saccadic movement can not be interpreted in the same way as fixations are, considering interest and attention towards the target of gaze.

Our ignorance about the exact relation between saccade properties and visual perception are of subordinate nature considering our rather coarse temporal resolution of gaze measurement (sample rate of 50 Hz).

3.2.4.2 Algorithms for fixation detection

Salvucci & Goldberg (2000) compared different algorithms that detect fixations from raw gaze data.

- A velocity-threshold applied to the point-to-point velocity can be used to distinguish between fixations and saccades. The velocity would for example not exceed a certain value during fixations but be higher than a threshold value during saccades.
- A dispersion-based algorithm uses a moving window. Consecutive data points are checked for the dispersion of their coordinates. If for a minimal duration of a sequence of data points a threshold for dispersion is not exceeded, this is considered as a fixation. The window is extended onto consecutive data points, incorporating them in the detected fixation, until a maximum threshold of dispersion is exceeded. Then the algorithm is reinitialized and the window is moved to the following data points.
- A two-state Hidden Markov model with velocity distributions for saccades and fixations as states is another possibility of a detection algorithm. For state changes a low transition probability should be chosen and a high probability of remaining in the current state. The HMM can learn the respective parameters from a previously labeled corpus.
- Another algorithm uses areas of interest, representing information units in the visual field, for fixation identification. Gaze points falling into specified rectangular target areas are considered as fixation points or as saccade points otherwise. Consecutive gaze points are grouped into a fixation if the duration of the sequence exceeds a temporal threshold. Otherwise they are considered as part of a saccade crossing the area and discarded.

Salvucci & Goldberg (2000) conclude that the Hidden Markov model and the dispersion-based algorithm provide both accurate and robust fixation identification, whereas the implementation of the Hidden Markov model is more intense. The detection using a velocity-threshold is the easiest to implement, but not very robust. They do not recommend the use of the algorithm using areas of interest the performance of which they describe as rather poor.

3.2.4.3 Choice of algorithm for fixation detection

With respect to the evaluation of the different algorithms by Salvucci & Goldberg (2000), we decided that for our purposes and under the given constraints a dispersion based approach for fixation detection was the most feasible. It needs no training compared to HMMs and is nevertheless equally robust but much simpler to implement. The rather low frequency of 50 Hz of the Tobii® eye trackers is not in favor of a velocity based approach. The algorithm using predefined areas of interest has lower performance and is not apt for our purposes, as our areas of interest cannot be attributed to fixed position in the video frame.

The dispersion-based approach exploits the fact that gaze points that belong to the same fixation are very close to each other, which is closely associated to the velocity properties of fixations and saccades. Consecutive gaze points are grouped into a fixation until a threshold is reached. This threshold is defined as a maximal distance between any two gaze points inside this interval. The algorithm starts with a defined minimum number of consecutive gaze points. If in this minimal interval the threshold is already exceeded, the interval is not considered as a fixation and the algorithm will move one step further in time. The first gaze point of the preceding interval will be dropped out and the next gaze point at the end incorporated into the current interval. When an interval meets the described criterion, it is considered as a fixation. The interval will be extended onto the following gaze points, until the threshold is exceeded.

3.2.4.4 Specification of parameters

The detected fixations depend partly on the choice of parameters for the initialization of the detection algorithm. The detection of fixations is stable to a certain extend and minor

alterations of parameters would produce only small variations of fixation properties. Fixations may be joined into longer fixations or split into several shorter fixations. The main information of target location and total duration of time dedicated to a target would however not substantially be altered. For the sake of the statistic evaluation of the duration of fixation this should however be kept in mind.

For the processing of our data, we defined a maximal window size of 10x10 pixels at a resolution of 320 to 240 pixels, and a minimum duration of 100ms which corresponds to 5 consecutive gaze points. Visual inspection revealed that the fixation detection algorithm delivers coherent results. Figure 46 shows a sample of gaze data and the detected fixations. No increased appearance of gaps or isolated gaze points between fixations can be observed, that would indicate a too conservative detection. The results are quite appropriate for the purpose of the current investigation.

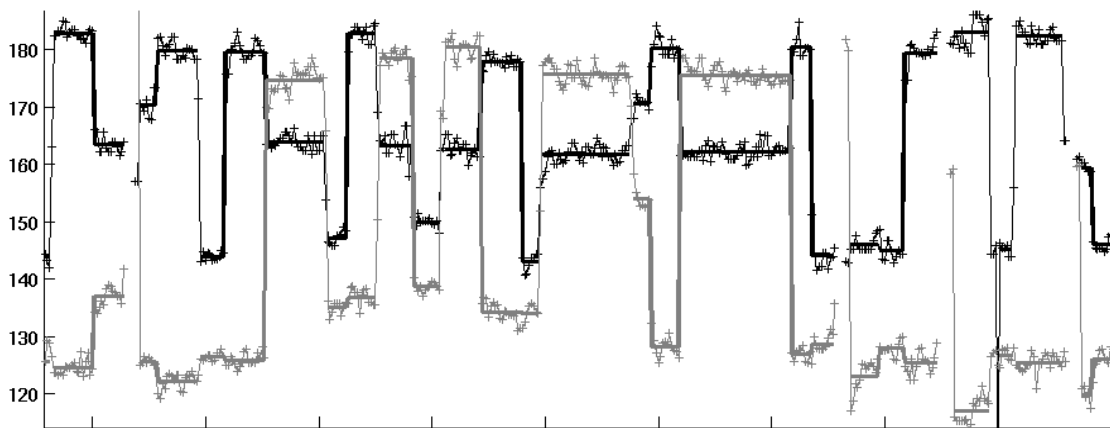


Figure 46: Extract from the graphical representation of the gaze direction of our target subject (as respondent in interaction 8). The x-coordinates of gaze direction are represented in black, y-coordinates in gray. Crosses mark gaze points in 20 ms steps on the time axes (abscissa) and in pixels on the ordinate whereas the gaze targets on the screen are represented at a resolution of 320x240 pixels. The horizontal bold lines represent detected fixations with the means of the x- and y-coordinates of the gaze points that belong to respective same fixation.

3.2.5 Assignment of regions of interest

For statistical analysis of the relation between fixation target and course of conversation, we distinguish different regions of interest. In the following, we detail the regions we distinguish and how we assign fixations to these regions.

3.2.5.1 Distinction of regions of interest

Once the gaze data is segmented into fixations, we assign these fixations to potential targets. These targets are regions on the face that are of special interest in human visual perception. We denominate these as ‘*regions of interest*’ (ROI) and distinguish the right and left eye, the mouth and the remaining parts of the face. We classify fixations that fall on other regions of the scene in the category ‘*else*’. Furthermore time spans, during which no fixations are detected are listed in the category ‘*none*’. Hence, we distinguish the labels *right eye*, *left eye*, *mouth*, *face*, *else* and *none* for the categorization of fixations.

3.2.5.2 Relating fixation to regions of interest

In a first attempt to assign gaze targets to regions of interest, we defined the ROI in every frame of the video. We tracked automatically the eyes and the mouth as anchor points, around which we calculated the ROI anew for every frame. Then the coordinates of the current

fixation were compared to these to check if they fell into one of the ROI. This was not satisfactory. Gaze often was directed towards areas that subjectively could be assigned to an ROI with high certainty, but did not fall into the predefined zone covering the respective ROI. Accumulations of fixations may for instance occur slightly below the eyes, but not fall into the zone defined around the eye. Therefore, manual assignment of ROI a posteriori was performed.

To assign the detected fixations to ROIs, all detected fixations are projected on a reference image. To eliminate microsaccades and other minor movements that occur during fixations, the mean of the concerned coordinates is chosen as representative value. This value is projected onto the reference image. All the fixations that occur during the part of the interaction that will be analyzed are displayed as asterisks on the same reference image. Their sizes correspond to their durations. Differently colored lines indicate the directions and magnitudes of the preceding and subsequent saccades. This information helps the experimenter to decide which fixations to enclose within the accumulation of fixations that will be assigned to a certain ROI. For this assignment, four points are defined around the accumulation to define an ellipse enclosing it. According to the visualization of this ellipse on the image, the assignment can either be accepted or modified until an optimal choice is made. The fixations are then tagged with a marker indicating the corresponding ROI for later analysis. In the reference image this is represented with a corresponding color of the asterisks.

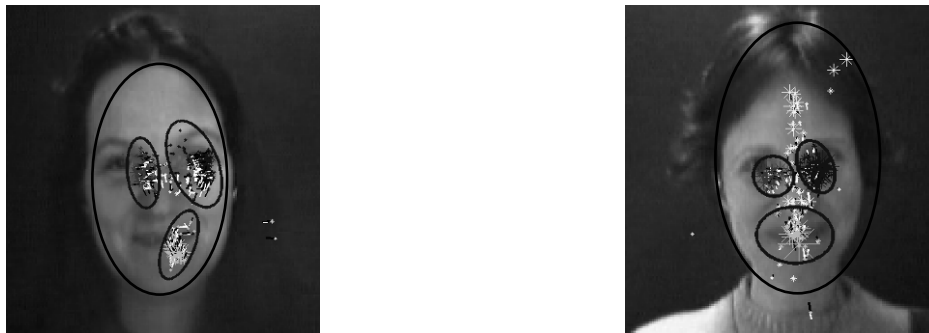


Figure 47: Assignment of fixations to regions of interest. The ellipses are determined by hand to delineate the fixations assigned to the ROI mouth, right eye, left eye and face. The size of the asterisks is proportional to the duration of fixations. Left: fixations of our target speaker on face of subject, Right: the interlocutor's data.

3.2.5.3 Tracking regions of interest

In order to assign fixations to ROI, localization of the anchor points of ROI should be performed. These anchor points are tracked automatically with an in-house software tool. The tool allows defining a search window around each track point together with the local appearance of the point. The tracking can be interrupted or reinitiated at any frame, in order to execute manual corrections or to relocate the search frame and track points. This is used to bridge events that mask the track points temporarily and to continue automatic tracking afterwards.

For the recordings discussed in this report, the two eyes and the mouth have been tracked. From these three reference points and the three corresponding points on a reference image, a bilinear transformation matrix is calculated that is used to project back the measured gaze data onto a single frame. For the representation of fixations on the target image, every fixation is represented by its respective mean coordinates that are transformed according to the transformation matrix.

For the purpose of demonstration and visual verification, videos are generated, in which the current gaze target of the interlocutor is displayed. As the frequency of gaze data is twice as high as the frame rate, there are two gaze targets per frame.

3.2.6 Detection of blinks

We also label blinks that are facial features closely related to eye movement. Intervals between fixations, during which no valid gaze data could be measured, are considered as potential blinks by the algorithm we use for automatic detection of blinks. According to this algorithm the duration of these intervals should lie between 20ms and 240ms and the fixations before and after potential blinks should have a minimum length of 160ms. These criteria produce a satisfactory automatic detection of blinks given a sampling rate of 50Hz which is of course rather coarse for eye gaze in this context. The automatic detection is followed by manual verification and correction to assure reliable labeling of blinks.

3.2.7 Segmentation of interaction

For the analysis of the experiments, we segmented the discourse according to the activity and the mental states of the subjects. With an in-house software tool, based on the energy and zero-crossing contours, the intervals of speech activity are automatically detected in the audio stream. As the two subjects are recorded on separate channels, the detected speech activity can directly be associated with each subject. With the use of headphones for the experiment, crosstalk could be prohibited and the utterances of the subjects could successfully be separated on the two channels of the stereo recording. We could do so since the microphones did not capture the other person's speech emitted from the earphones.

3.2.7.1 Automatic detection of cognitive states

Due to the imposed repetitive structure of the recorded interactions, the detected intervals of speech activity are used to derive the sequence of cognitive states (CS). Once the respective role of subjects is known, an automatic segmentation of the discourse is possible, except for the CS *pre-phonation*.

Table 7 explains how the CS can be determined from the possible combinations of speech state, and how the CS of the initiator are related to the CS of the respondent. When both, initiator and respondent are not speaking in the current interval, the CS is derived from their speech activity in the preceding interval. Only in a few cases, the automatically labeled sequences have to be modified. Intervals that could not be assigned with these CS and therefore had to be declared manually as *else*, were for instance periods of laughter.

When an interval of speech activity is detected on an audio channel, the CS can reliably be set to *speaking* for the corresponding subject, and to *listening* for the other subject. In the following the CS depend on the role. In the role initiator, *speaking* may be followed by a CS *waiting*, if the respondent does not immediately repeat the sentence. The time the respondent passes to prepare the utterance in mind is classified as *thinking* and is eventually followed by an interval of *pre-phonation*. When the respondent finally utters the sentence, the initiator correspondingly is *listening*. Once the respondent terminated the utterance, the initiator will read the next sentence to utter which is labeled as *reading*.

When acting as respondent, *speaking* is followed by *waiting*, which is the interval while the initiator is reading the next sentence. Then follows *listening*, while the initiator speaks, eventually *thinking*, to remember and organize the just heard sentence before *speaking* again to repeat it. As already mentioned, *speaking* is usually preceded by an interval of *pre-phonation*.

To ease the manual annotation of *pre-phonation* a default interval is added during automatic segmentation before every interval of CS *speaking*. This is only done if the state preceding *speaking* exceeds the sum of the default interval duration chosen for *pre-phonation* and the predefined minimal interval duration. We chose a default duration of 400ms for the CS *pre-phonation* and 80ms as the minimum duration for intervals. The default intervals of *pre-phonation* set during the automatic segmentation are later corrected by hand and the duration is adjusted or the interval deleted if no *pre-phonation* can be observed.

A typical sequence of cognitive states as resulting from the segmentation of the recorded interactions is given in table 8.

Table 7: Possible combinations of speech activity for the initiator and respondent, and the consecutive classification of the respective intervals (1 represents speaking). If both subjects are not speaking in the current interval (speech state 0 at t), the CS is determined from the speech activity in the preceding intervals (speech state t-1).

speech state at t-1		speech state at t		classification of CS at t	
initiator	respondent	initiator	respondent	initiator	respondent
0	0	0	0	?	?
0	1	0	0	read + pre-ph.	waiting
1	0	0	0	waiting	thinking + pre-ph.
1	1	0	0	?	?
?	?	0	1	listening	speaking
?	?	1	0	speaking	listening
?	?	1	1	speaking	speaking

Table 8: Typical sequence of cognitive states resulting from the segmentation of the recorded interactions.

initiator	<i>speaking</i>	<i>waiting</i>		<i>listening</i>	<i>reading</i>	<i>pre-phon.</i>	<i>speaking</i>	...
respondent	<i>listening</i>	<i>thinking</i>	<i>pre-phon.</i>	<i>speaking</i>	<i>waiting</i>		<i>listening</i>	...

3.2.7.2 Manual verification of automatic segmentation

For the manual verification and correction of the automatically generated segmentation, we chose the ELAN® multimedia annotation software (Hellwig & Uytvanck (2004)). This is a tool for the annotation of video and audio resources (see <http://www.lamp-mpi.eu/tools/elan/elan-description>). It has been developed at the Max Planck Institute for Psycholinguistics at Nijmegen in the Netherlands.

Annotations can be created on multiple layers, called tiers that can be interconnected hierarchically. The transcription is stored in an XML format. ELAN® delegates media playback to an existing media framework, like Windows Media Player, QuickTime or JMF (Java Media Framework). As a result a wide variety of audio and video formats is supported and high performance media playback can be achieved.

From the automatic segmentation as described above, three tiers per subject are created: cognitive state, fixation target and blink. To inform about the role, a supplementary tier is displayed that informs about the boundaries and category of role. Another two tiers of labels were generated for the purpose of later analysis. They are built from the fixation targets that are further subdivided taking the boundaries of the CS-labels into account. Therefore, the fixation labels are split at the boundaries of cognitive state intervals. This eases estimation of temporal relations between gaze targets and cognitive state.

Originally, ELAN® is designed for manual annotation of resources from subjective observation. As in our experiments we extracted most of the annotation boundaries

automatically from the measured gaze and speech data, we developed special scripts to create the annotation tiers automatically from this data.

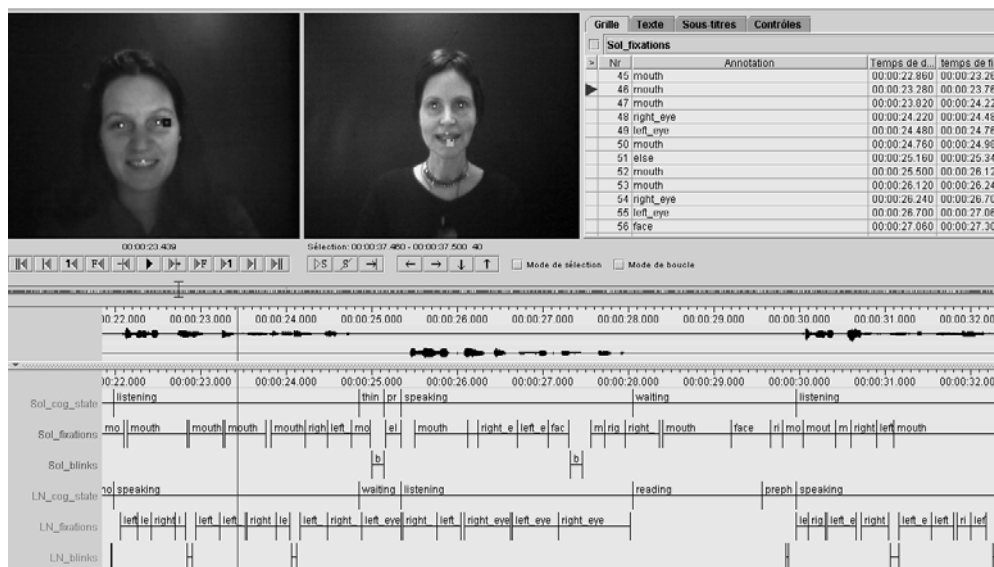


Figure 48: Multimedia annotation software ELAN® (Hellwig & Uytvanck (2004)) as used for manual verification and correction of automatic segmentation. Top: video images with superimposed fixation targets. Middle: audio signal. Bottom: annotation tiers for cognitive state, fixation and blink.

3.3 STATISTICAL ANALYSIS

Two approaches were considered for the analysis of the measured data. On the one hand, we seek for relations between the inner states and activities of a subject during the interaction and gaze behavior. This is an endogenous control of gaze by the inner mental processes of a subject and can be considered as a direct conscious or unconscious manifestation of cognitive states. On the other hand, we seek for relations between the gaze behavior of one subject, and the gaze behavior of the other. Especially in the case of eye contact, we expected to observe either gaze avoidance or a search for contact. As this would mean an external influence, we consider this as exogenously controlled gaze behavior. We investigated the gaze of a subject around events of eye-directed gaze of the other as a response to external excitation.

3.3.1 Endogenous gaze control

In the context of our experiments, we consider the influence of the cognitive state and role of a subject on its own gaze behavior as endogenous influence on gaze behavior. For statistical analysis, we thus considered duration of fixation, fixation time, probability of fixation and blink frequency as variables that may vary in dependence of CS and role.

The following discussions mainly concern data measured on our target subject. In order to reach statistical significance, we combine all the CS-intervals of our target subject over all interactions. As we recorded nine such interactions, with ten sentences per role, there are about 90 instances of every combination of CS and role. Peculiarities in the interaction originating from the personality of the interlocutor or the attitude of the target subject that may vary over time should be leveled out by this measure and the general behavior of the target subject should emerge from this data. If data measured on the interacting subjects is discussed, this will be explicitly mentioned in the text.

3.3.1.1 Fixation time

The degree to which a ROI on the face of the interlocutor is of interest as a source of information as well as its importance in the context of social norms in dialogue should influence the amount of gaze that is dedicated to this ROI during a given CS. To be able to compare the amount of gaze independently of the duration of the CS, we put the fixation time measured during a CS in relation to its duration to obtain the percentage value.

This results in six values for every instance of a CS that inform about the proportion of fixation time dedicated to the ROI *face*, *right eye*, *left eye*, *mouth* or *else*, as well as the proportion of time that no fixations were detected (*none*).

Descriptive statistics

To allow for detailed visual inspection of the acquired data, we displayed the observations of fixation time, separated for CS and role as box plots (see Figure 49). For this purpose, we used the Matlab function *boxplot* that generates graphical representations of the repartition of data.

The horizontal lines of the box represent the lower quartile, the median and the upper quartile values. The whiskers, displayed as vertical dashed lines, extend from each end of the box to the most extreme values within 1.5 times the interquartile range. Values beyond the ends of the whiskers are treated as outliers and are represented by crosses.

The notches in the box to the sides of the median value display the uncertainty of the estimation of the median for comparison between boxes. Notches of boxes that do not overlap indicate a significant difference between the respective medians at the 5% significance level. The significance level is based on the assumption of a normal distribution. The comparisons of medians are supposed to be a robust estimate. They function as a visual hypothesis test, analogous to the *t*-test used for means.

The box plots of fixation time for the CS during the role *initiator* are in the left column of diagrams, those for the CS during the role *respondent* are in the right column of Figure 49. Below each diagram the corresponding CS is indicated, as well as the number *n* of instances observed for the respective combination of CS and role.

The box plot diagrams show a strong variation of fixation time over the different ROI. According to the visual comparison of the notches of the box plots inside a given combination of CS and role, there are significant differences between several of the distributions of fixation time. This indicates that gaze is not evenly distributed over the different possible targets. The fact that the individual distributions vary over the different combinations of CS and role, suggests that these have an influence on the observed repartitions. The relation between CS, role and repartition of fixation time over the possible ROI is of great interest and the object of the following statistical analysis.

For an overview over the repartition of fixation time and for direct visual comparison between CS and roles, Figure 50 displays the mean values calculated over all interactions in 3D bar plots. In contrast to the box plots that allow a detailed visual comparison of the distributions, the three dimensional bars plot present only the means of these distributions.

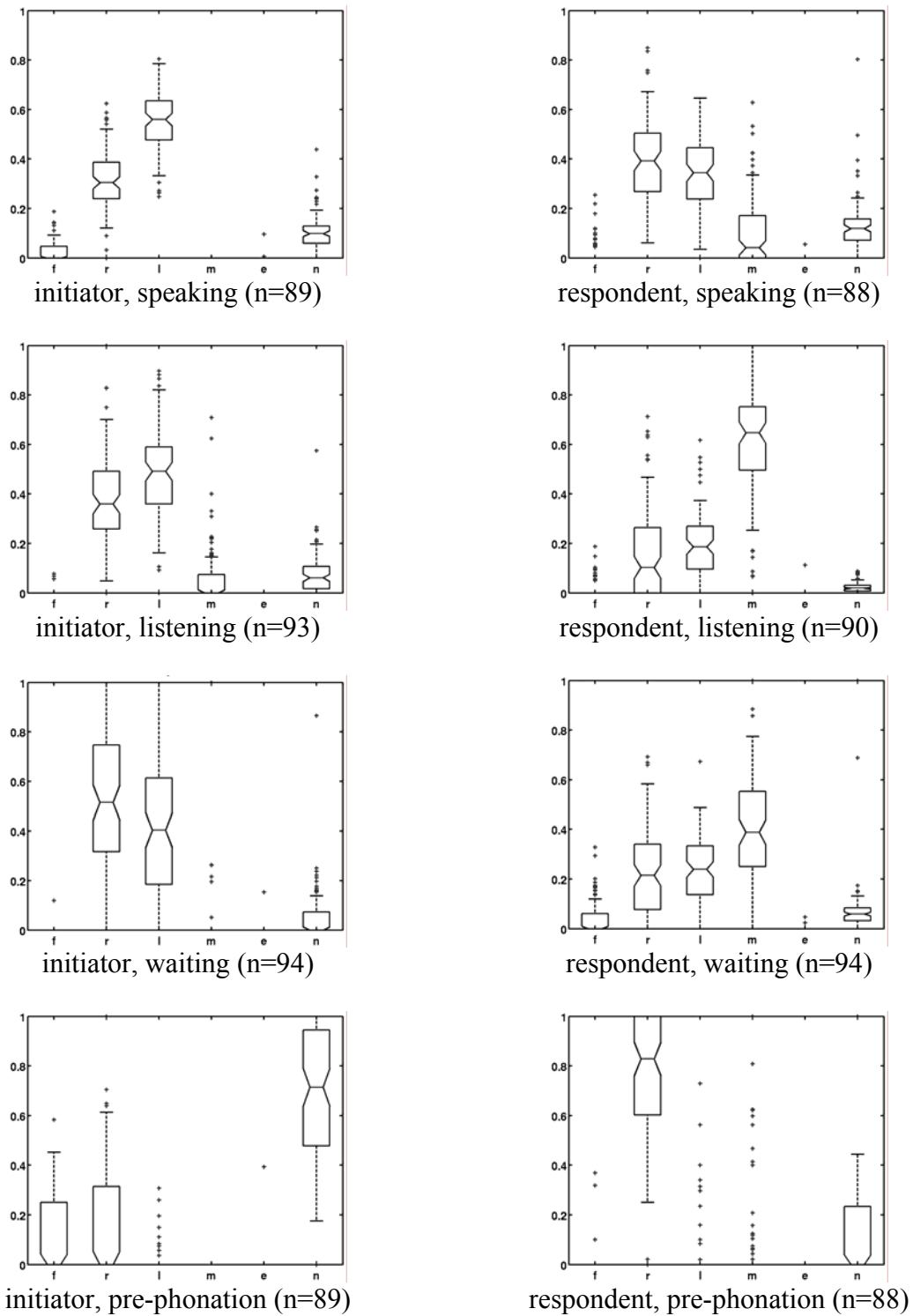


Figure 49: Box-plots of repartition of fixation time during an instance of a cognitive state during a given role. The box-plots are calculated from the proportion of time that was dedicated to a ROI during the instances of a cognitive state. Here the ROI face, right eye, left eye and mouth that were taken into account for the statistical test are listed.

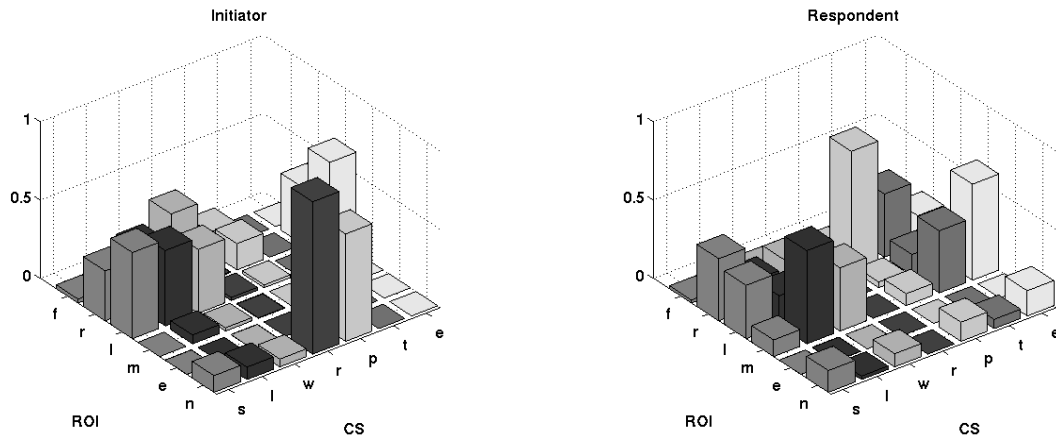


Figure 50: The diagrams show a 3D representation of the repartition of mean fixation time dedicated to the different ROI in %/100. The bars represent mean values of the proportion of time that was dedicated to a ROI during the instances of a cognitive state. A detailed representation of the repartition of data is given in Figure 49. Abscise: speaking, listening, waiting, reading, pre-phonation, thinking else; ordinate: face, right eye, left eye, mouth, else (represented by initial letter).

Analytic statistics

As an objective measure, we conducted a series of analysis of variance tests to compare the observed distributions of fixation time. As we are interested in its repartition over the possible ROI as a sort of fingerprint of the combinations of CS and role, we used the *MANOVA* function that Matlab proposes to compare multivariate means.

MANOVA is an extension of ANOVA simultaneously taking into account multiple variables as coordinates of a multidimensional vector. The observations specify points in the multidimensional space defined by the variables as axes. MANOVA tests whether different groups, built from combinations of points, build separate accumulations in this space. Therefore, orthogonal axes are calculated as linear combinations from the original axis-vectors to achieve maximum separation between the groups. The first axis describes the strongest separation between the groups, with declining level of separation for the following axes. MANOVA tests the null hypothesis that the mean vectors of the groups lie on various dimensions and indicates the number of dimensions in which the means are estimated to be significantly different. For all tested dimensions, a corresponding *p*-value is calculated. The degrees of freedom are specified as degrees of freedom of the within groups sum of squares (*dfW*), degrees of freedom for the between group sum of squares (*dfB*) and degrees of freedom for the total sum of squares (*dfT*). The maximal dimensionality of separation for a given number of groups, is either limited by the number of groups, or by the number of components of the multivariate factor. Four groups for instance can be maximally separated on three dimensions, as four points are necessary to define a 3-dimensional space. Independent of these constraints, there is a certain dependency between the measured values that may influence the dimensionality of separation. The components of the multivariate factor are determined as proportions of the durations of a CS as the whole and therefore add up to one for every instance of a CS. They are however not absolutely dependant, as only four of the six values are used.

For the MANOVA tests, we took only those CS into account, that occur in both roles. These are *speaking*, *listening*, *waiting* and *pre-phonation*. The factors considered for the analysis are hence, CS and role.

As the depending multivariate variable of the analysis, we consider ROI, distinguishing *face*, *right eye*, *left eye* and *mouth* as its components. The ROI *else* and *none* are not considered, as they occur very irregularly and therefore do not contribute to distinguish the influence of the different factors.

- **MANOVA over CS, independent of role**

Testing the four groups built by the CS *speaking, listening, waiting* and *pre-phonation*, independent of role, MANOVA determined that the group means are significantly separated on $d = 3$ dimensions. The corresponding p-values and degrees of freedom are:

- $p_1 \leq 0.01$ $p_2 = \leq 0.01$ $p_3 = 0.0231$
- $dfW = 721$ $dfB = 3$ $dfT = 724$

- **MANOVA over all combinations of CS and role**

A MANOVA comparing all combinations of CS and role resulted in 3-dimensional separation of groups ($d = 3$):

- $p_1 \leq 0.01$ $p_2 = \leq 0.01$ $p_3 \leq 0.01$ $p_4 = 0.1393$
- $dfW = 717$ $dfB = 7$ $dfT = 724$

A representation of the data on the three orthogonal axes of maximum separation of groups, calculated from all combinations of CS and role, is displayed in Figure 51. The ellipses represent the respective combinations of CS and role. Some of them overlap strongly, but some are clearly separated and further apart.

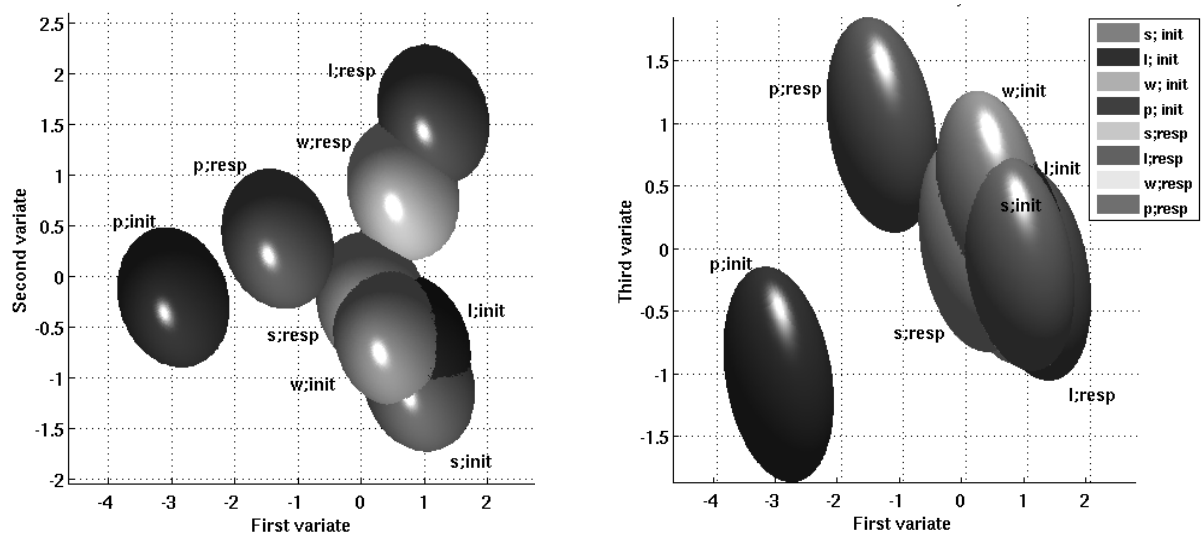


Figure 51: Projection on the first discrimination planes of the groups formed by the combinations of CS and role as indicated in the diagrams with the respective abbreviations. The groups are represented as a projection on the first two main axes in the diagram to the left, and on the first and third axis in the diagram to the right.

- **MANOVA over CS, separated for role**

A MANOVA comparing the CS separately for each role resulted also in 3-dimensional separation of groups ($d = 3$) for both roles.

Results for role *initiator*:

- $p_1 \leq 0.01$ $p_2 = 9.03 \cdot 10^{-12}$ $p_3 = 1.6 \cdot 10^{-4}$
- $dfW = 361$ $dfB = 3$ $dfT = 364$

Results for role *respondent*:

- $p_1 \leq 0.01$ $p_2 \leq 0.01$ $p_3 = 5.2 \cdot 10^{-4}$
- $dfW = 356$ $dfB = 3$ $dfT = 359$

- **Pair wise MANOVA over combinations of CS and role**

Pair wise comparisons of CS are given in Table 9. The upper right triangle of the table lists comparisons between different CS of the same role *initiator*. The lower left triangle of the table lists comparisons between different CS of the same role *respondent*. In the diagonal, the results from the comparison of groups with identical CS but different role are listed.

As only two groups are compared, the maximal dimensionality is one. In all cases, the MANOVA resulted in significant separation of groups in one dimension, indicating that compared CS are significantly different. The corresponding *p*-values are given in the table, along with the number of observations and number of variables taken into account. When comparing *speaking* to *pre-phonation* in the role *initiator*, only 3 variables were taken into account as the variable *mouth* only contained zeros. The respective degrees of freedom arise from the given number *n* of samples. There are *n*-2 degrees of freedom of the within groups sum of squares *dfW*, *n*-1 degrees of freedom for the total sum of squares *dfT*, and one degree of freedom for the between group sum of squares *dfB*.

Table 9: Pair wise comparison of multivariate means with MANOVA. The upper and lower part of the table list the respective pair wise comparisons of CS within the same role. The cells in the diagonal list the results of the comparison of the two roles of a same CS. The cells indicate the *p*-value, indicating the significance of separation between the groups and the number *n* of samples included in the analysis (observations times variables taken into account).

		initiator			
		speaking	listening	waiting	pre-phonation
respondent	speaking	< 0.01; 177· 4	$7.83 * 10^{-11}$; 182· 4	$7.62 * 10^{-13}$; 183· 4	< 0.01; 178· 3
	listening	< 0.01; 178· 4	< 0.01; 183· 4	0.0017; 187· 4	< 0.01; 182· 4
	waiting	< 0.01; 182· 4	$4.83 * 10^{-8}$; 184· 4	< 0.01; 188· 4	< 0.01; 183· 4
	pre-phonation	< 0.01; 176· 4	< 0.01; 178· 4	< 0.01; 182· 4	< 0.01; 177· 4

Discussion

For the analysis, we segmented the dialogue into intervals of different cognitive states and assigned them to one of the two roles that subjects have in the interaction. From the significance of the separation of gaze data between these groups, we conclude that the categorization we applied to the data is coherent. The factors have significant influence on the repartition of fixation time over the different ROI.

We conducted a control test in order to confirm the significance of test results. An arbitrary grouping of data did not result in a significant separation of means when tested with a MANOVA. For each of the possible combinations of the CS *speaking* and *listening* with the two roles, we built two arbitrary subgroups. None of these groups showed significantly different means ($d = 0$ in all tested cases). This result confirms that variations within CS are not significant, while variations between CS are.

Inspecting the diagrams, the eyes emerge as a very dominant target of fixations. This confirms the observations of other researchers, who reported the eyes and the mouth as salient targets (see section 1.3).

In our data, only in the role *respondent* the mouth becomes a frequent target, especially while *listening*. Most probably, this is in order to decode additional information from lip reading. This partly confirms Lansing & McConkie (1999), who report that subjects show a stronger tendency to gaze at the mouth when attentive to segmental information compared to cases where subjects are mainly attentive to prosodic information (see section 1.3.4, page 42).

There may however also be a social component to the increased fixations towards the mouth, for instance as a signal of attention and interest.

A remarkable peak of fixations towards the *right eye* appears in the role *respondent* for the CS *pre-phonation*. *Pre-phonation* intervals are usually very short, and occur during the upward gaze shift between *reading* and *speaking*. It seems as if for our target subject the *right eye* was the default target when starting to look at a face after a period of averted gaze. It might also be related to her taking the turn, which she does at the same instance.

It would be interesting to put our results in relation to the findings of Kendon (1967) about the proportions of gaze directed away or at the interlocutor. Kendon reported them to vary between speaking and listening. In our data however, the gaze of subjects was almost always directed to regions of the other person's face. In cases where the amount of gaze not directed to the face of the interlocutor is high, this is due to the fact that the subject has to read instructions, in order to know the next sentence to utter. Otherwise, the percentage of gaze that is directed towards the interlocutor's face is close to 100%. This is obviously due to the nature of the interaction that is subject to certain restrictions and very different from the kind of interaction that Kendon (1967) recorded for his analysis. A direct comparison of the results is therefore not possible.

With respect to the modeling of gaze behavior for the animation of an embodied conversational agent, our results are a strong argument to consider the inner states of subjects as endogenous influence on the control of gaze behavior. If humans do expect a relation between the gaze behavior and the activity of an interlocutor in conversational interaction, the variation of gaze direction with respect to role and CS is important to generate meaningful and believable gaze behavior. For the same purpose, we claim that a random variation of angles of gaze orientation assuming the interlocutor straight ahead is not sufficient. For coherent gaze animation the exact knowledge of the position of an interlocutor as well as of the eyes are relevant to signal consciousness of environment.

3.3.1.2 Probability of fixations over ROI

In the bar plot diagrams in Figure 50, only means of proportional fixation times are displayed, calculated over all instances of the same CS of the nine analyzed interactions. They do not unveil information about the distribution of fixation time over the individual instances of CS. In this respect, the representation in box plots (see Figure 49) gives more information. To get better insight of how important the individual ROI are in a given combination of CS and role, we examined the probability of ROI being fixated at least once during every instance of a given CS.

If two ROI are equally important and fulfill similar functions for a given CS, gaze may either be directed to one or the other. This may result in a probability of 0.5 for both. If both were important individually, this would result in probabilities close to one for both of them. The mean value of fixation time may be identical in both cases and therefore would not inform about such an individual importance of ROI. Figure 52 displays these probabilities in two bar-plots, one for each role.

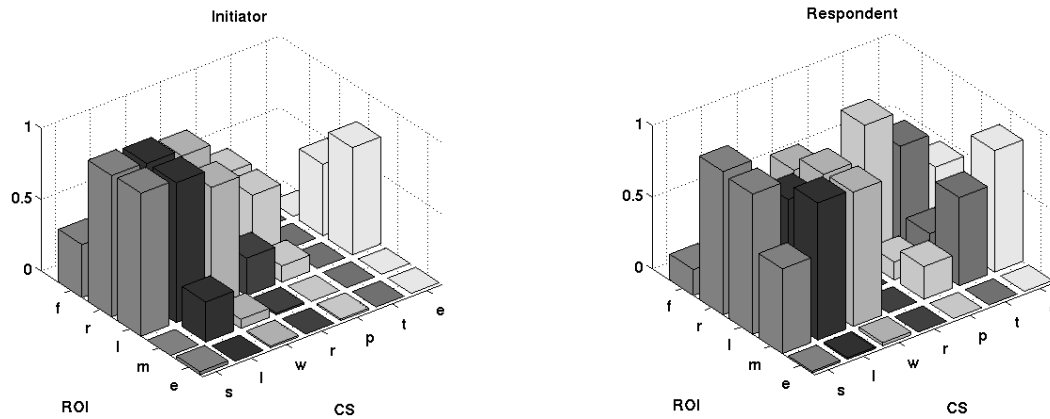


Figure 52: The diagrams show a 3D representation of the probabilities that a ROI is fixated at least once during a CS in %/100. Abscise: speaking, listening, waiting, reading, pre-phonation, thinking else; ordinate: face, right eye, left eye, mouth, else (represented by initial letter).

In all CS, the two eyes are highly probable targets of fixations. Exceptions are instances of CS *thinking* while *initiator* and *reading* while *respondent* – these combinations are of no importance, as they do never occur.

Especially in the role *initiator*, the eyes are fixated at least once during nearly every instance of the CS *speaking*, *listening* and *waiting*. The difference produced by the role *respondent* is remarkable. In this case, *mouth* becomes a predominant target with a very high probability to be fixated during the CS *listening* and *waiting*. In the case of CS *speaking* it attains a probability of $p \approx 0.5$.

As already mentioned when discussing the bar plots of fixation time (see section above), the differences of fixation probability between the two roles are probably due to the different degree of acquaintance with the content of the utterance. When hearing the utterances as *respondent*, their content is new and unknown. The mouth is then an important target to acquire additional information from lip-reading and receives tendentially more attention. The *initiator* already knows the content and does not need such information, as it is only a repetition of the sentence she previously uttered. Similar to the interpretation of the fixation time data, the increased probability of fixations at the mouth confirms the findings of Lansing & McConkie (1999), who report a tendency to gaze at the mouth when subjects are attentive to segmental information.

While repeating the sentence, fixations to the mouth may be related to the process of remembering the previously heard sentence. When referencing objects, people tend to gaze to where they remember them. In this case, the mouth is the origin of the visual manifestation of the sentence to repeat and may be fixated for this reason. At the same time, the *initiator* may deliver information from mouthing when repeating the sentence quietly along with the *respondent*. We did however not observe a pronounced tendency of the subjects to do so as *initiators*.

Independent of the increased probability of gazes towards the mouth, the eyes are still important targets of fixations and are fixated with high probability. This may be due to the social aspect of signaling attention.

Remarkable is also the strong tendency to fixate the right eye during the CS *pre-phonation* as *respondent* that also emerges in Figure 50. It seems that this temporarily very restricted and in its visual manifestation very prominent dialogic event, is closely associated to a single ROI. It may be interesting to see if there are events in dialogue and interaction that have such a strong association to fixation targets.

Comparing the bar plot diagrams of fixation time and fixation probability, both give of course a similar statement about fixation targets. There are however quantitative differences. For the modeling of gaze behavior, we should find a solution that covers both aspects.

3.3.1.3 Duration of fixation

For the analysis of duration of fixation, we examined the same factors - CS and role. We consider the natural logarithm of the measured duration in order to obtain distributions that are closer to Gaussian distributions. The histogram of log-transformed durations of all fixations detected in the data (from all interactions of our target subject) shows a bimodal distribution (see Figure 53). The mean over all fixations calculated from the log-transformed data corresponds to the mean duration of 317ms.

The two peaks suggest the existence of two main categories of fixations with different distributions of durations. We consider that this depends on the fixated target. We thus compute distributions for the five ROI, as displayed in Figure 54. As there were few fixations to targets on the screen other than on face regions (ROI *else*), these are neglected in the following. For the other ROI, three groups can be distinguished from visual examination. An analysis of variance test shows significant differences between these groups (ANOVA: $df_{group} = 9$, $df_{total} = 2990$, $F = 39.71$, $p \leq 0.01$). Fixations to the face tend to be shorter, fixations to the mouth tend to be longer and fixations to the eyes lie in between. The one by one comparison with a multiple comparison test using *Tukey's honestly significant difference criterion* shows that these three groups are significantly different at $\alpha = 0.01$. The means and standard deviations of the distributions are indicated besides the dashed vertical line representing the mean in the figures. The duration corresponding to the mean in milliseconds is given in braces.

A further separation of fixations that distinguishes role does not add any further distinctions (ANOVA: $df_{group} = 4$, $df_{total} = 2990$, $F = 78.56$, $p \leq 0.01$). A one by one comparison shows the same differences between the three groups, significant at $\alpha = 0.01$. The mouth as a group however nearly disappears for the role *initiator* due to only few fixations (see Figure 55).

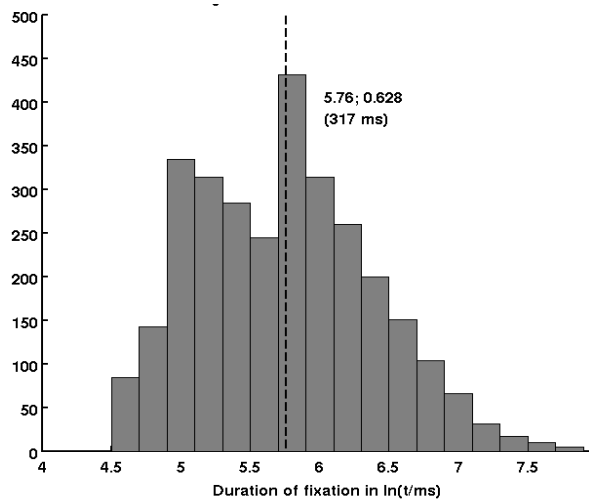


Figure 53: Histogram of the duration of all fixations of the reference subject of all interactions. To obtain a symmetrical distribution the natural logarithm of the values is taken. The dashed line indicates the mean, which is also given as value, along with the standard deviation and the duration in ms corresponding to the mean.

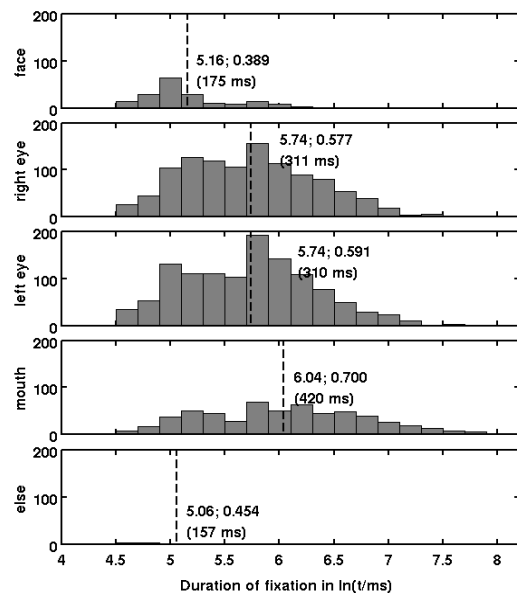


Figure 54: Same as Figure 53 but grouped for the CS during which the fixations were observed.

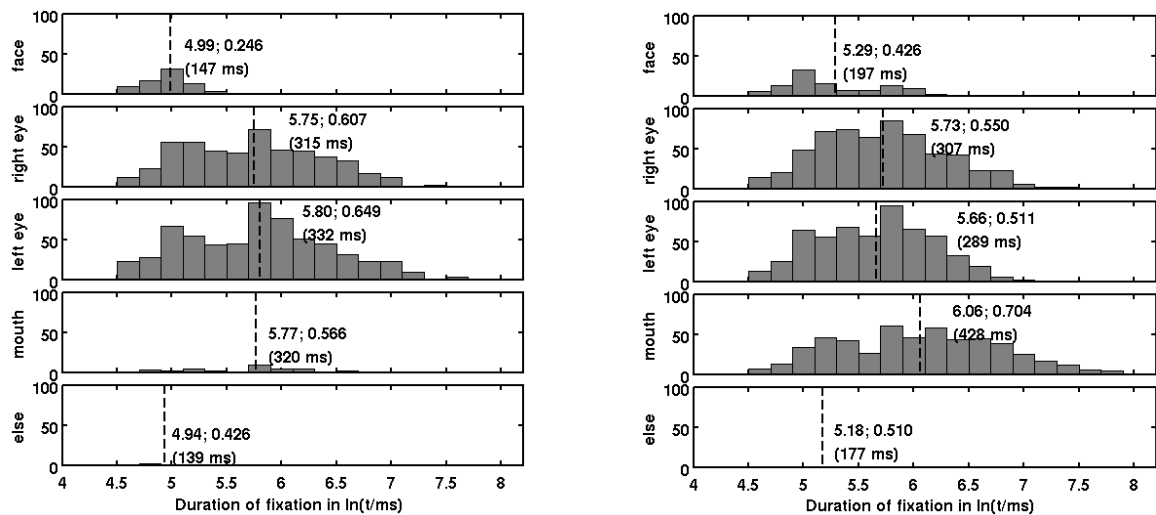


Figure 55: Same as Figure 54 but grouped for the CS and the role during which the fixations are observed. The left diagram shows the histogram for the role of *initiator*, the right diagram shows the histogram for the role of *respondent*.

The analysis reveals significant differences of duration of fixation depending on the target of fixations. The distributions of durations observed for fixations at either of the two eyes, are very similar and not significantly different. Fixations to the face however are in mean significantly shorter, and fixations to the mouth significantly longer. The generation of fixations with a gaze model should consider these differences.

The analysis we performed on the data did not reveal any explanation for the broader distributions of durations of fixations at the eyes and their slight secondary peaks. It would be interesting to know if there are factors that can explain these characteristics, which deserves further investigation. The hypothesis, that fixations of short duration are refixations of the same target could not be confirmed. When only considering the fixations that are directly preceded by a fixation to the same target, the obtained histograms show still the same tendencies.

3.3.1.4 Blinks

Blinking is not directly related to gaze orientation and should be considered as a separate facial movement. We chose to include it in this analysis, since it affects gaze perception and paces exploration of our field of vision. During the duration of a blink, the eyelids cover and replace the eyes as stimulus. They are therefore closely linked to the perception of gaze. Furthermore, blinks are reported to be influenced by mental conditions (see section 0, page 25). We consider that the states we distinguish may have a similar influence. The scenario and setup are entirely appropriate to the inclusion of the examination of blinks in our analysis.

Although certain parameters are known to influence blink rate, as for instance cognitive load, emotional state or fatigue, there to our knowledge is no empirical model nor precise theory about the occurrence of blinks on which our analysis can be grounded. We considered two different hypotheses about the occurrence of blinks for the further analysis. Blinks occur entirely at random. Then their number should be roughly proportional to the length of examined intervals. They might also occur at a regular frequency, at regular temporal intervals. Variations in the individual duration of inter-blink intervals may be influenced by cognitive states or special events. We analyzed the detected blinks according to these two assumptions.

Random occurrence of blinks

If blinks occurred at random, the mean frequency calculated for a certain interval, should approach the overall mean frequency, if the length of this interval approached the total duration of the analyzed sequence. In this case, the over all mean frequency is the expected frequency of blinking for the inspected interval. Dividing the number of blinks observed in this interval by its duration, we obtain the observed frequency.

We combined the intervals of the same CS separated for role over all interactions of our target subject. For such combinations of intervals, the mean frequency, calculated from the number of blinks concerned by these intervals and the corresponding total duration of these intervals, should also approach the over all mean frequency. We hypothesize that this should be true, independent of CS and role, if blinks occurred at random. To verify this hypothesis we compared the observed number of blinks to the number of blinks expected according to the over all mean frequency with a test, known as Pearson's *chi-square test* (χ^2) (see equation 1). For the comparison of observed and expected number of blink, we did not take into account the CS *reading* and *thinking*, since they do not occur equally over roles.

$$\chi_{n-1}^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

O_i = observed frequency
 E_i = expected frequency
 n = the number of possible outcomes of each event

(1)

$$\chi_{n-1}^2 = \sum_{i=1}^n \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

O_i = observed frequency
 E_i = expected frequency
 n = the number of possible outcomes of each event

(2)

In a first step we estimated whether interaction (interacting subject) is a factor of influence on the observed number of blinks. Therefore, we calculated the tables of contingency with the number of blinks observed during the different interactions (see Table 10). As in some CS mainly small frequencies are expected we applied *Yates' correction of continuity* (see equation 2). The test showed that only interaction 7 rejects the hypothesis that the numbers of blinks are independent of the interaction. This means that for our target subject the blink behavior is relatively constant over interactions and we did not consider interlocutor as a factor for further analysis.

The observations of blinks for the respective interactions already show a tendency that suggests an influence of CS and role. The intervals of CS *speaking* and *listening* are for instance of about the same duration and should therefore produce about the same numbers of blinks. This is however not the case. During *listening* there are strikingly less blinks. In contrast, there are a lot of blinks in the CS *pre-phonation* in spite of relatively short durations.

Table 10: Table of observed blinks during the different CS (*speaking, listening, waiting, pre-phonation*) in the role *initiator* and *respondent* over the different interactions of the target subject.

interlocutor	initiator				respondent			
	s	l	w	p	s	l	w	p
1	14	1	2	9	15	0	7	0
2	10	0	1	10	20	0	6	1
3	13	0	1	9	17	0	7	0
4	17	0	0	9	14	1	4	5
5	13	4	2	9	11	0	12	5
6	10	3	1	4	11	0	17	2
7	18	7	2	9	19	6	13	6
8	21	2	1	6	15	0	3	2
9	10	4	0	6	12	0	11	4

To test if blinks appear at random, we counted the number of blinks observed for each type of CS over all interactions of our target subject. These are the observed number of blinks, as listed in Table 11 (left). From the corresponding total duration of combinations of CS and role and the over all mean frequency, we calculated the expected number of blinks (Table 11, right). We did so independently of role as well as separating between roles. The observed numbers of blinks are compared to the expected ones with Pearson's *chi-square test* (χ^2).

As we included four different CS in the analysis, there are three degrees of freedom, which corresponds to the critical value $\chi^2 = 7.82$ for alpha = 0.05 and $\chi^2 = 11.35$ for alpha = 0.01. The obtained values of $\chi^2 = 321.1$, $\chi^2 = 163.8$ and $\chi^2 = 443.4$ are far beyond these critical values. Therefore, the hypothesis of random occurrence of blinks is rejected.

Table 11: Table of observed number of blinks summed over all interactions of the target subject for the different CS and corresponding number expected according to the mean frequency of blink. The values of χ^2 are calculated separating role, and independent of role.

CS	observed number of blinks				expected according to over all mean blink frequency				χ^2
	s	l	w	p	s	l	w	p	
Initiator	126	21	10	71	71.3	72	18.7	13.7	321.1
Respondent	134	7	80	25	63.7	72	91	9.3	163.8
All	260	28	90	96	135.0	144	109.4	23.0	443.4

A representation of the blink rate observed during the interactions of our target subject is displayed in Figure 56. A frequency in the common sense cannot be determined when distinguishing CS, as often only one blink occurs per instance of a CS. There are not enough inter-blink intervals inside a CS to compute a reliable mean frequency. Intervals of *pre-phonation* for example are very short and usually only one blink occurs within each CS. Therefore, the displayed blink rate is calculated as the total number of blinks over the total sum of durations of a CS for a given interaction and role. It is remarkable to see the high blink rate for *pre-phonation* in spite of the very short durations of the intervals, especially for the role *initiator*.

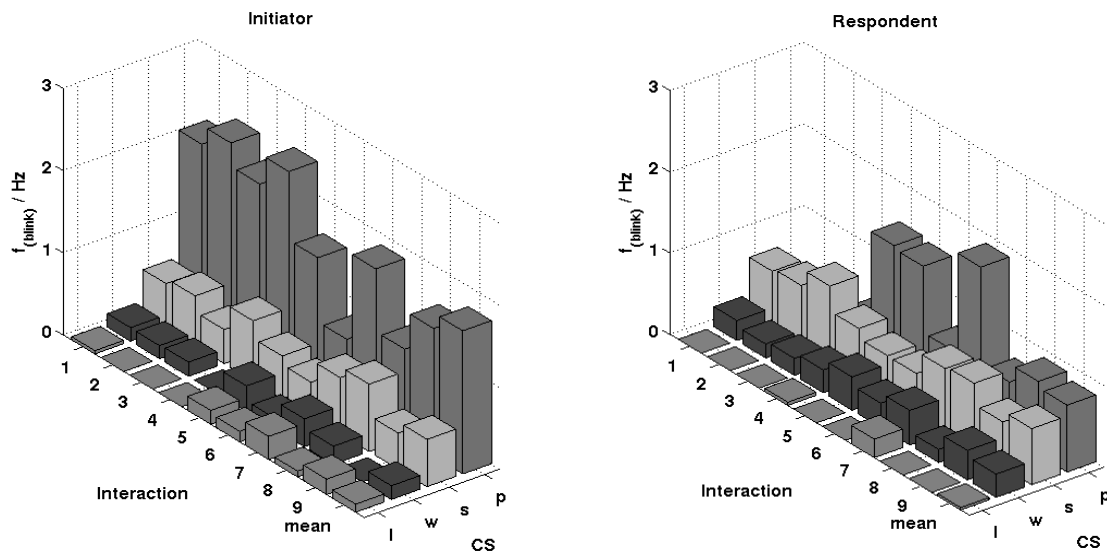


Figure 56: Blink rate, calculated as total number of blinks per sum of duration of CS for a given interaction (1 ... 9) and role (left: initiator, right: respondent). Abscise: listening, waiting, speaking, pre-phonation (represented by intial letter)

To generate a representation of the blink data not related to the respective length of CS, we calculate the probabilities that at least one blink appears during an instance of a CS separately for all interactions and separated for role (see Table 12, *initiator* and Table 13, *respondent*). The tables show that the probabilities are very high for the CS *speaking* and *pre-phonation*. In spite of the fact that the intervals of *speaking* and *listening* are very similar in length, the probability that a blink appears while *listening* is much lower, especially in the role *respondent*. There is also a difference between roles for the CS *waiting*, for which blinks are more frequent in the role *respondent*. This may be due to the difference in length of CS. In the case of *initiator*, an interval of CS *waiting* corresponds to the CS *thinking* of the *respondent* which is of very short duration. This corresponds to the short interval when the interlocutor (being *respondent*) is preparing in mind the previously heard sentence before repeating it. In the other case, in the role *respondent*, the CS *waiting* appears while the interlocutor (being *initiator*) is reading the next sentence to utter, which takes relatively more time.

Table 12: Probabilities that at least one blink appears during an instance of a CS, over all interactions, for the role *initiator*. Furthermore, the mean probabilities are calculated from the probabilities of the individual interactions in the role *initiator*. Note that there is no CS *thinking* in the role *initiator*.

interaction	speaking	listening	waiting	reading	pre-phonation	(thinking)	else
1	0.9	0.07	0.1	0	0.9	-	-
2	0.7	0	0.1	0.3	1.0	-	-
3	0.8	0	0.1	0	0.9	-	-
4	1.0	0	0	0.1	0.9	-	0
5	0.9	0.4	0.2	0	0.9	-	0
6	0.8	0.3	0.1	0.2	0.4	-	-
7	1.0	0.6	0.2	0.1	0.9	-	0
8	1.0	0.2	0.1	0	0.6	-	-
9	1.0	0.4	0	0.3	0.7	-	0
mean	0.9	0.2	0.1	0.1	0.8	-	-

Table 13: Probabilities that at least one blink appears during an instance of a CS, over all interactions, for the role *respondent*. Furthermore, the mean probabilities are calculated from the probabilities of the individual interactions in the role *respondent*. Note that there is no CS reading in the role *respondent*.

interaction	speaking	listening	waiting	(reading)	pre- phonation	thinking	else
1	0.8	0	0.6	-	0	0	0.2
2	1.0	0	0.5	-	0.1	0	-
3	1.0	0	0.6	-	0	0	0
4	1.0	0.1	0.4	-	0.5	0	-
5	0.8	0	0.6	-	0.5	0	-
6	0.9	0	0.8	-	0.2	0	-
7	1.0	0.5	0.8	-	0.6	0	-
8	1.0	0	0.3	-	0.3	0	-
9	0.9	0	0.9	-	0.4	0	-
mean	0.9	0.1	0.6	-	0.3	0	-

Constant frequency of blinks

If blinks were a regularly occurring event deferring more or less from an otherwise constant frequency, the interval between two consecutive blinks should follow a Gaussian distribution centered at this frequency. This is however not the case for measured inter blink intervals (see Figure 57). In the following we investigate if there is a relation between inter blink interval duration and CS or role.

Testing consistency of data over interactions

In a first step, we compared the blink frequencies observed for our target subject during the different interactions, to control if interaction should be considered as factor for the analysis. Table 14 lists the standard deviation and the mean of the inter-blink interval duration as well as the corresponding mean blink frequency for the different interactions. The last column indicates the values calculated over all interactions.

An ANOVA showed significant differences of interval length between the different interactions ($df_{group} = 8$, $df_{total} = 467$, $F = 3.54$, $p = 0.0005$). A post hoc multiple comparison test using *Tukey's honestly significant difference criterion* showed that only for interactions 6 and 7 there are significantly different distributions of inter-blink intervals, but that no other pairs are significantly different. In fact, in interaction 7 our target subject shows a very high blink frequency compared to the other interactions. The high blink rate in interaction 7 is also visible in other graphical representations, especially in the case of CS *listening* (see Table 10, Figure 56, Table 12 and Table 13).

The different characteristics of blink observed in interaction 7 compared to the other interactions may be due to the special friendly relations between this subject and our target subject. Visual inspection of the video recordings did not suggest any other explanation for the increased blink frequency, such as for instance irritation of the eyes. As only the pair wise comparison of interaction 6 and 7 gives reason to assume an influence of interaction on blink, we did not consider interaction as a factor for the analysis.

Table 14: Mean and standard deviation of inter-blink interval duration and corresponding mean blink frequency as measured for the different interaction of our target subject duration

Interaction	1	2	3	4	5	6	7	8	9	all
mean /s	3.62	3.55	2.99	2.74	2.67	3.55	2	2.73	3.06	2.92
stdv / s	2.68	2.63	2.27	2.16	1.71	2.02	1.08	2.17	1.45	2.07
f / Hz	0.28	0.28	0.34	0.37	0.38	0.28	0.5	0.37	0.33	0.34

I - Investigating influence of CS and role on inter-blink interval duration

Figure 57 shows histograms of the distribution of all observed inter-blink intervals as measured in milliseconds and transformed to the natural logarithm. Especially in the transformed data, two separated peaks emerge. The peaks might be related to different combinations of CS and role.

To verify the hypothesis that blinks occur at a regular frequency, we compared all inter-blink intervals to the mean inter-blink interval duration of 2.92s and the standard deviation of 2.07s, calculated from the data gathering all interactions of our target subject. We defined a criterion determining whether we consider a blink as delayed or inhibited by a given instance of a CS considering these values. Only intervals of CS that exceed a minimum duration of 400ms were taken into account. Other intervals were considered as too short to have a major influence on the occurrence of blink. In Figure 58 the deviations of these interval durations from the mean, are displayed in histograms, distinguishing role and CS.

If the interval since the last blink exceeds the standard deviation for the current CS and no blink occurs during this CS, it is considered as inhibiting blinks (see Figure 59 for example). This applies also to subsequent CS if they exceed the minimum duration of 400ms and still no blink occurs. The numbers of CS that inhibited blink according to this criterion are listed in Table 15.

Observations:

- The deviation from the mean duration of an inter-blink interval appearing during or overlapping a CS *speaking*, is mainly negative. Speaking shows a strong tendency to shorten inter-blink intervals compared to the mean.
- There is a difference between roles for the CS *speaking*. When *respondent*, there is a peak of a short positive delay, which is not found in the role *initiator*.
- *Listening* tends to increase inter-blink intervals.
- Intervals between blinks that occur in the context of *waiting*, show a deviation from the mean that is equally distributed around zero, whereas they occur only rarely in the role *initiator*. This is probably influenced by the irregular appearance of the CS *waiting* and their relatively short durations in this role, as already stated above.

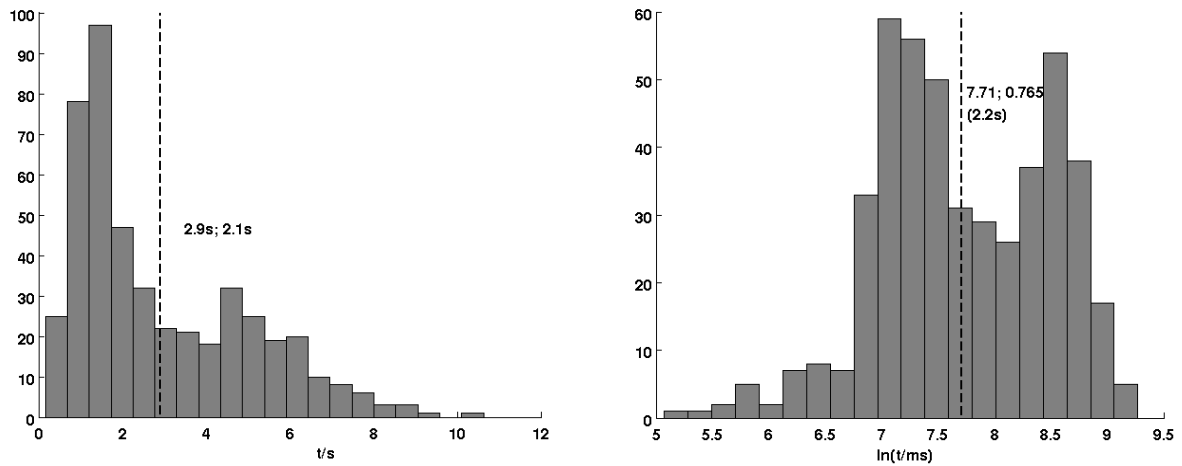


Figure 57: Histogram of all inter-blink intervals measured during the interactions of our target subject, represented as measured in milliseconds (left) and after transformation to the natural logarithm (right). The dashed line indicates the mean, given also as a figure in the diagram along with the standard deviation and the corresponding mean in milliseconds in the case of the logarithmic representation.

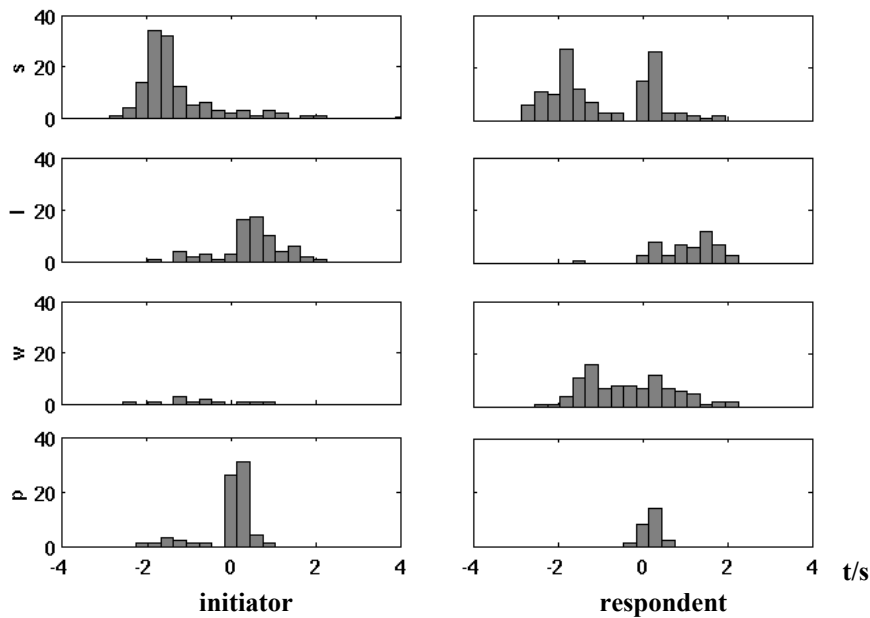


Figure 58: Deviation of inter-blink duration from mean, measured during the interactions of our target subject, separated for CS and role (*initiator* left; *respondent* right). Only the CS *speaking*, *listening*, *waiting* and *pre-phonation* that are common to both roles and appear with a minimum frequency are considered.

Table 15 : Number of blinks inhibited by the different CS during a given role due to a criterion taking into account the mean and standard deviation of inter-blink interval duration. The first pair of lines is calculated taking all inter-blink intervals into account. For the determination of the second pair of lines, the mean and standard deviation have been recalculated excluding the intervals from the above lines, and blinks that are very close to the start or end of a cognitive state.

		speaking	listening	waiting	reading	pre-phon.	thinking	else
I	Initiator	0	5	2	34	2	0	0
	Respondent	2	25	1	0	5	3	3
II	Initiator	1	18	3	53	5	0	0
	Respondent	2	53	6	0	7	4	5

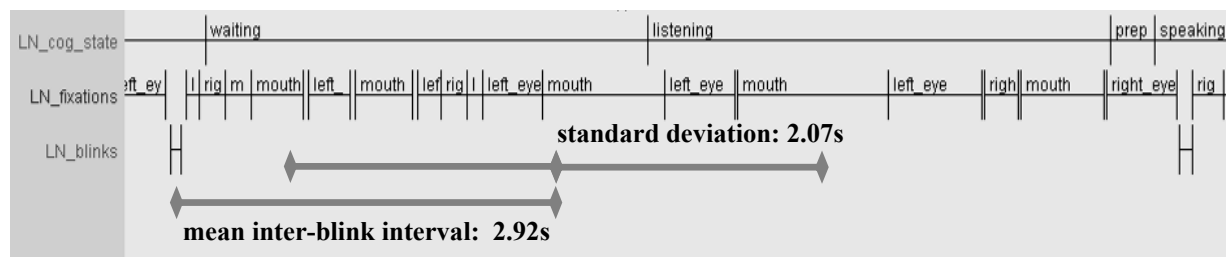


Figure 59: Screen capture from the graphical interface of the annotation tool ELAN® showing an example of an inter-blink interval strongly delayed relative to the mean. The concerned CS *listening* is hence classified as inhibiting blink.

II - Excluding blinks close to boundaries of CS

We considered the possibility that the end or the beginning of a CS may trigger blink and excluded these cases in a second analysis. Blinks that are closer than 300ms to the boundary of a CS are now excluded from the evaluation along with the concerned inter-blink intervals and the mean and standard deviation are computed again. The resulting data is displayed in Figure 60 in histograms of inter-blink intervals, and in Figure 61 in histogram of the deviation from the mean interval duration.

After the filtering of data according to this criterion, the second peak in the CS *speaking* when *respondent* disappears. This indicates that the positive delay that appears as a distinct peak in Figure 58 (upper row, right), is not due to a blink interval occurring inside an interval of CS. These blinks are preceded by states that do not favor blinks. A blink may then occur, as soon as a CS not impeding blink succeeds, following the retained impulse to blink. This is obviously the case for CS *speaking* when *respondent* and produces the peak of short positive delay. This suggests the assumption that there is an influence of CS on blink competing with physiological needs.

The eight groups of deviations from the mean inter-blink interval duration, displayed as histograms in Figure 61, show significant differences when tested for with an ANOVA, ($df = 7, F = 39.36, p \leq 0.01$). The results of a multiple comparison test using *Tukey's honestly significant difference criterion* show several differences between the individual groups. The CS *speaking* are (independent of role) different from the CS *waiting* and *listening* at $p \leq 0.01$. *Waiting* in the role *respondent* is significantly different from *listening* in the role *initiator* at $p \leq 0.01$, and at $p \leq 0.05$ from *listening* in the role *respondent*. These statistics corroborate the influence of CS and role on the occurrence of blink.

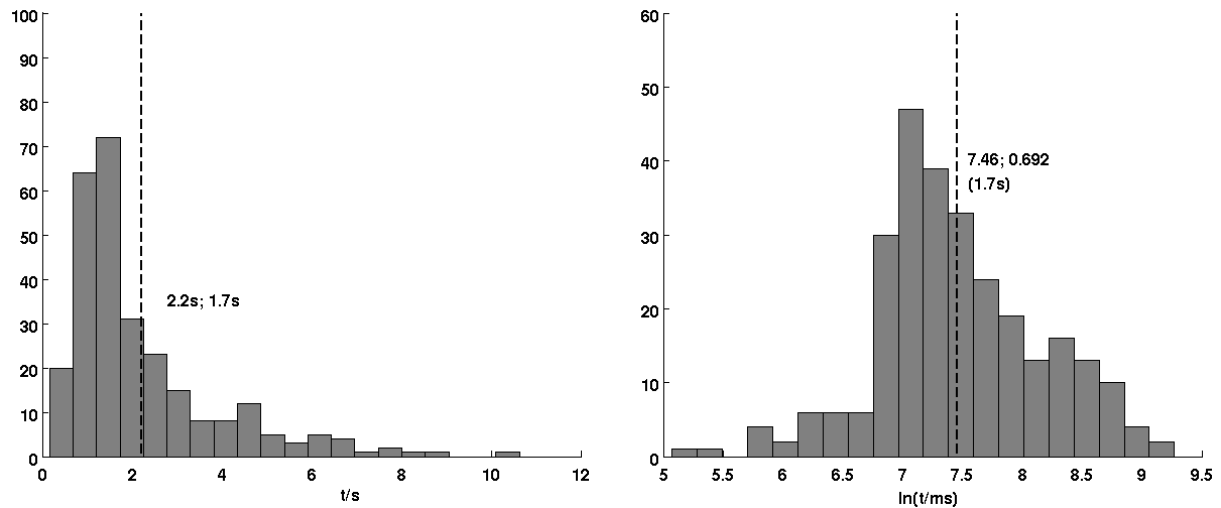


Figure 60: Histogram of all inter-blink intervals measured during the interactions of our target subject that can not be associated with the start or end of a CS. They are represented as measured in milliseconds (left) and after transformation to the natural logarithm (right). The dashed line indicates the mean, given also as a figure in the diagram along with the standard deviation and the corresponding mean in milliseconds in the case of the logarithmic representation.

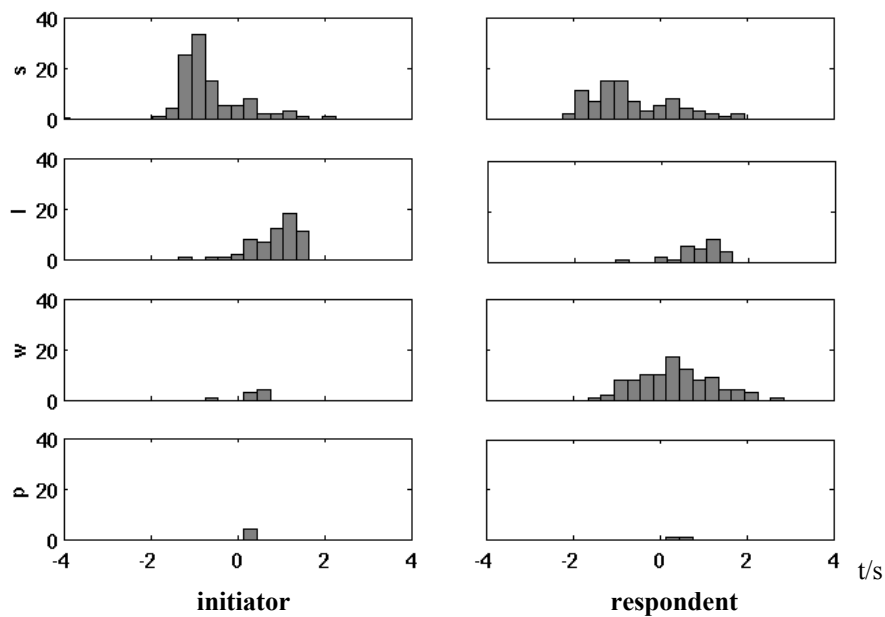


Figure 61: Deviation of inter-blink duration from mean, measured during the interactions of our target subject, separated for CS and role. Only the CS *speaking*, *listening*, *waiting* and *pre-phonation* that are common to both roles and appear with a minimum frequency are considered. The left diagram displays all measured intervals. In the right diagram, intervals that can be associated with the start or end of a CS are excluded. Data show that listening tends to slow down blink rate whereas speaking speeds up the blink rate.

Discussion

The analysis of blink gives strong evidence that the occurrence of blink depends on CS and role. With *chi-square tests* we showed that the occurrence of blink follows patterns that are consistent over the different interactions of our target subject and do not correspond to random occurrence. The probability of occurrence of blink varies over CS and role. It depends more on the interaction of CS and role of an interval, than on its duration.

The investigation whether blinks occur at a regular frequency showed that there are distinct deviations from the mean inter-blink interval duration. Whether the interval is longer or

shorter than the mean depends strongly on the nature of the concerned CS. For cases of inter-blink intervals exceeding the mean duration by more than the standard deviation, we considered that there is an inhibitory influence. Such inhibitory influence is also closely related to the interaction of CS and role.

Generalizing these observations, there is a tendency to avoid blinks with increased attention. Such instances are for example the CS *listening*, especially in the role *respondent*, and *reading* (when *initiator*).

When listening to a sentence in the role *initiator*, the content is already known to the listener, as she previously uttered the same sentence. The attention is therefore lower than when listening to sentence with unfamiliar content in the role *respondent*. This may explain the differences observed for the CS *listening* between the two roles. We take this as an argument to assume that an increase of attention results in a decrease of blink rate in a continuous way, and that it is not a binary function of CS.

The blinks occurring regularly during *pre-phonation* cannot be related in the same way to the intensity of attention. Most probably these are related to turn taking. The increased occurrence of blink during *pre-phonation* resembles to the aversion of gaze at the beginning of a turn. Kendon (1967) considered such aversion of gaze as a possible strategy to insist on the turn, preventing competing signals from the interlocutor.

These blinks may however also be a protective reflex triggered by major head movements (Evinger, Manning, Pellegrini, Basso, Powers & Sibony (1994)). *Pre-phonation* occurs especially after the CS *reading* in the role *respondent*. While reading, the subject looks down on the paper notes and moves her head up to face the interlocutor when starting to speak.

The considerations about the blink behavior of our target subject are probably of general validity. The observations made for one of the subjects for instance, that shows extraordinarily few blinks, confirms this assumption. She produces only 14 blinks while being initiator, of which 11 during the CS *pre-phonation*, and two while *speaking* (1 during CS *else* (laughing)). When respondent, she produced only 6 blinks: 4 while *speaking*, 1 during *pre-phonation* and one during CS *else* (interruption for reflection during speaking). Although she produced extremely few blinks, they occur exclusively during cognitive states of reduced attention or during the CS *pre-phonation*. The regular occurrence during *pre-phonation* in spite of her strong tendency to avoid blink, is an argument to assume a communicative function of these blinks.

3.3.2 Exogenous control

3.3.2.1 Response to eye-directed gaze

In the previous section, we investigated the influence of the inner states of our target subject on her gaze behavior. In addition, we are interested to know if there is also an influence of exogenous parameters, originating especially from the interlocutor's gaze behavior. In this context, eye-directed gaze is of particular interest, since it is expected to trigger the strongest reactions. Analysis should reveal if eye-directed gaze produces reflex responses of the observer, observable as recurrent gaze behavior. We hypothesize for example that subjects either search for eye contact or on the contrary tend to avoid eye contact.

With respect to our target subject, we consider two possible signal-response relations. Either the target subject influences the gaze behavior of her interlocutors, or the interlocutors' eye-directed gaze influences hers. For both cases, we calculate mean signals and responses.

In both cases, the mean signal is calculated from all eye-directed fixations. The corresponding mean responses are calculated from the gaze of the respective interlocutors, separated for ROI. Every eye-directed fixation not directly preceded by a fixation towards the same target is taken into account for this analysis. We also discard any eye-directed fixation not preceded by

at least 240ms of non-eye-directed gaze. Around the onset of the signal as temporal point of reference, the gaze was charted as either directed (assigned value 1) or not (assigned value 0) towards a given ROI. For the fixations considered as response, we calculated separated curves for each of the ROI *right eye*, *left eye*, and *mouth*. Calculations are made in 20ms steps according to the resolution of gaze monitoring with the eye trackers.

In order to obtain the mean responses the respective charts are summed up grouped for ROI and role. The sum is then divided by their respective number of occurrence. The resulting curves are displayed in Figure 62 for the gaze of the target subject as response to the interlocutors' gaze as signal.

The mean signal of eye-directed fixations is clearly distinguishable as a peak, starting at $t=0$ on the time axis, which represents the onset of the eye-directed fixations. The curves representing the responses are rather flat and do not show any prominent movement of the curve that would suggest a direct relation between the gaze of interacting subjects. We repeated the procedure inverting the signal-response relation between the subjects. Figure 63 shows the curves measured when the eye-directed gaze of the target subject is considered as signal, with the gaze of the interlocutors as response. In this case, too, there are no prominent movements of the curve that would allow establishing a relation between the gaze patterns of the interacting subjects.

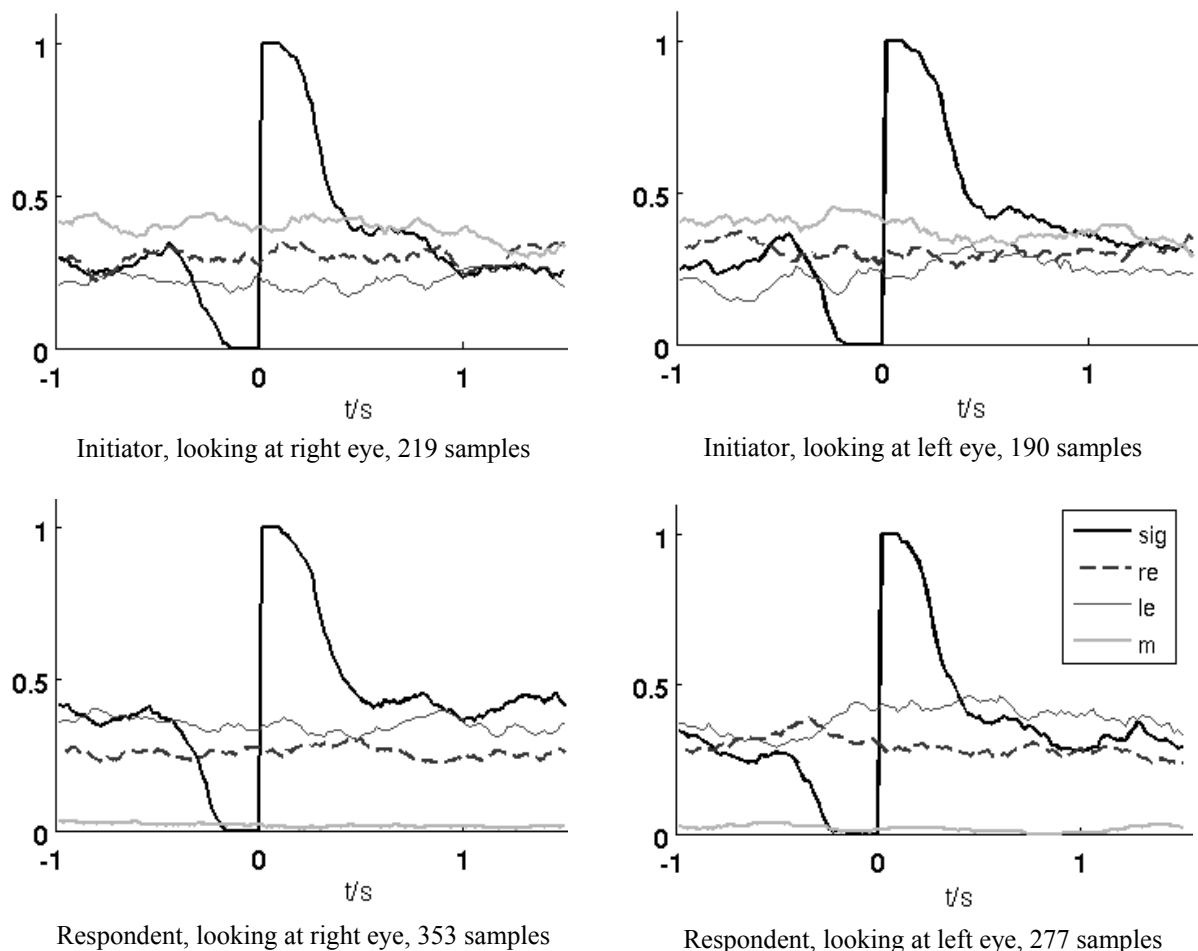


Figure 62 : Mean response of target subject to eye-directed gaze of interacting subjects as signal (continuous line). For every instance that a subject looked at the right or left eye of the target subject, the gaze direction of the latter was noted in 20ms steps as either directed (1) or not (0) towards a ROI (- mouth; -- right eye; - left eye). From the observations of all intervals taken into account the mean response was calculated, as the sum divided by the number of instances. The peak represents the mean signal of eye-directed gaze of the subjects.

The present diagrams confirms neither the hypotheses that subjects search for eye contact nor that they tend to avoid eye contact. The possibility that there are such relations can however not definitely be excluded. These may exist and depend on further factors that we do not take into account in the current analysis. More sophisticated strategies of analysis may be able to detect such relations as for instance the algorithm proposed in Magnusson (2000) for the detection of temporal patterns of larger scope and complexity. Another option is that the scenario and the restrictions made for the selection of subjects do not favor such mutual influence on gaze patterns.

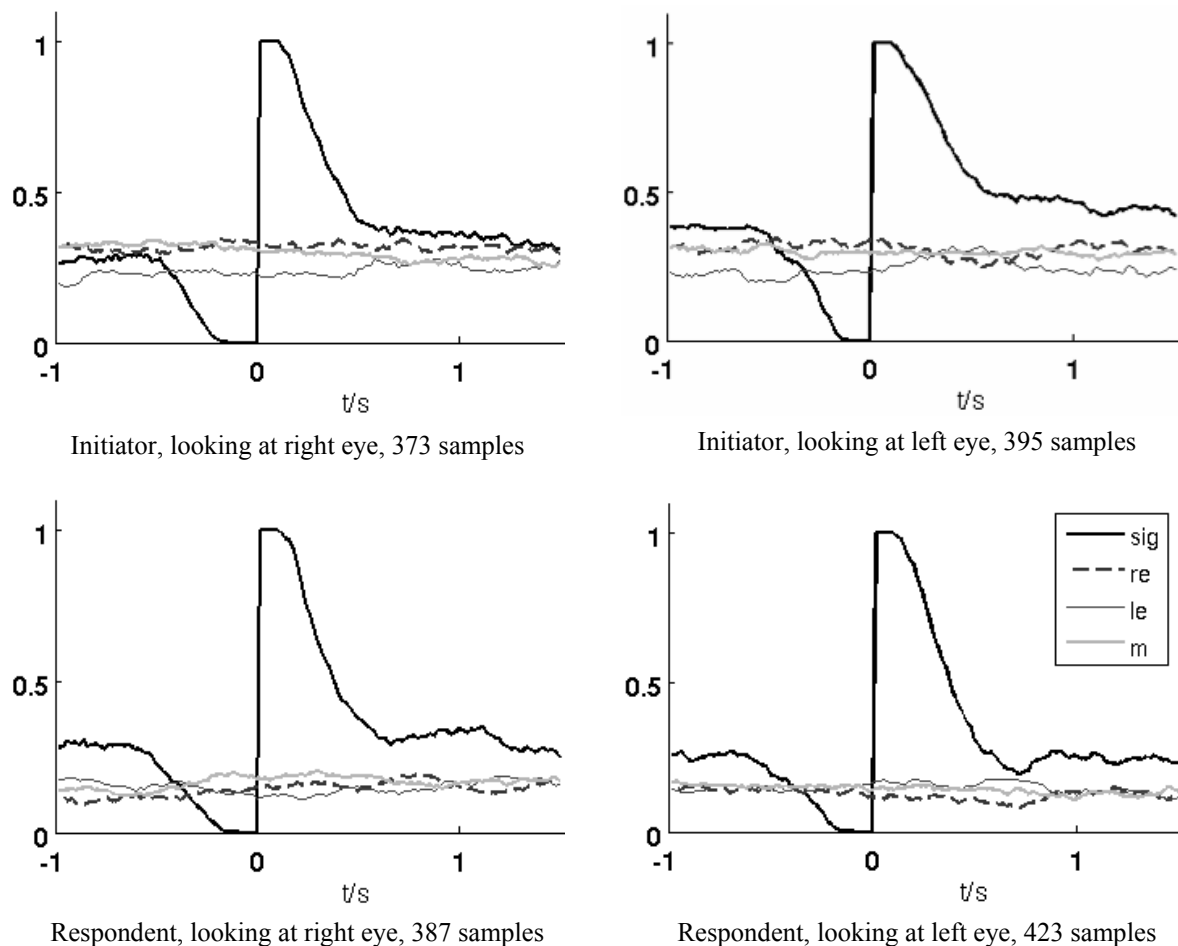


Figure 63: Mean response of interacting subjects to eye-directed gaze of target subject as signal (continuous line). For every instance that the target subject looked at the right or left eye of the interlocutor, the gaze direction of the latter was noted in 20ms steps as either directed (1) or not (0) towards a ROI (- mouth; -- right eye; - left eye). From the observations of all intervals taken into account the mean response was calculated, as the sum divided by the number instances. The peak represents the mean signal of eye-directed gaze of the target subject.

3.3.2.2 Mutual Gaze

In order to get detailed information about how eye contact varies depending on CS and role we calculated the percentage of time that mutual gaze occurs in dependence of these variables. Therefore, the durations of intervals during which both subjects directed the gaze to the each other's eyes are added up separated for CS and role and divided by the total sum of duration of the respective intervals. Figure 64 represents the results for the nine interactions in three dimensional bar plots separated for role. The mean calculated over all interactions is given in addition. We chose the CS *speaking*, *listening*, *waiting* and *pre-phonation* as these occur in both roles. As a matter of course, the two subjects cannot be in the same CS at the

same time. When the target subject is in on CS and role, the interlocutors are in the respectively corresponding CS and role, for instance *listening* while the target subject is *speaking*. The CS and role indicated in the diagrams therefore refer to the target subject and function only as reference intervals concerning the interlocutor.

The diagrams show major differences depending on the interaction as well as role. Comparing between roles, there is a general decrease of mutual gaze when the target subject is *respondent*. This cannot be explained by the role as such, as the gaze of both subjects has to be considered for the calculation of mutual gaze. There is therefore necessarily one subject in the role *initiator* while the other is in the role *respondent*. The observed differences probably depend on the interaction of role and personality of the subject, but cannot be explained by role alone.

In order to understand the differences in amount of mutual gaze observable in Figure 64, we further inspected the data stepwise. A prerequisite to the occurrence of mutual gaze is that there is eye-directed gaze. We inspected the eye-directed gaze of the target subject as well as of the interlocutors, to know if either of them favored mutual gaze more than the other. Figure 65 shows the percentage of eye-directed gaze during the CS *speaking*, *listening*, *waiting* and *pre-phonation* over the nine interactions, separated for role. Figure 66 shows the corresponding values of the respective interlocutors. It must be noted that only for the CS *speaking* and *listening* CS of the target subject can directly be related to CS of the interlocutor. As an example, the CS *speaking* in the role *initiator* of one subject corresponds to the CS *listening* in the role *respondent* of the interlocutor. For the other CS there is no such direct relation. The CS *waiting* in the role *respondent* for instance corresponds to the CS *reading* of the interlocutor, during which mutual gaze is very rare as the reading subject directs the gaze to the list of sentences. To enable a direct comparison of events, the intervals as considered in Figure 66 and their separation for CS and role corresponds to the segmentation and labeling of the target subject's data.

The bar plots show that the percentage of eye-directed gaze of the target subject is very consistent over interactions. It is relatively high and shows variations between CS. For all CS but *pre-phonation*, there is less eye-directed gaze in the role *respondent*. This is due to the fact that the CS *pre-phonation* in the role *initiator* is directly following the CS *reading* and the eyes are still directed towards the sentences that the subject just read.

The eye-directed gaze of the interlocutor subjects varies strongly between subjects (see Figure 66), which is not astonishing as the gaze behavior is known to vary strongly with personality. Some of the subjects show very few eye-directed gaze. This allows of course only for little amount of mutual gaze. Comparing Figure 64 and Figure 66, this relation is very obvious and explains the strong variations observed in mutual gaze. Especially the interactions 1, 8 and 9 clearly illustrate this relation.

In Figure 67 and Figure 68, we put mutual gaze in relation to eye-directed gaze. In Figure 67 the reference is eye-directed gaze of the target subject, in Figure 68 it is the eye-directed gaze of the interlocutors. Whereas this relation strongly varies over interactions in Figure 67, it is rather constant and high in Figure 68.

We conclude that our target subject has a distinct tendency to direct gaze towards the eyes of her interlocutor, and is open to mutual gaze. This tendency is quite consistent over interactions and hence independent of the interlocutor. If she fails to establish eye contact, this mainly depends on the interlocutor. Figure 67 and Figure 68 that show the mutual gaze in relation to eye-directed gaze confirm this conclusion. The high values in Figure 68 indicate, that in most cases when the interlocutor subjects look at the eyes and is ready for eye contact, very often the target subject answers this equally with eye-directed gaze, which results in mutual gaze.

The detailed investigation of the relation between eye-directed gaze and the occurrence of mutual gaze suggest a high readiness for mutual gaze of our target subject. This informs in the first place about the individual behavior of our target subject. It serves however also as a indication, that the gaze of the interlocutors has no major impact on the gaze behavior of our target subject. Independent of their amount of eye-directed gaze, she maintains her tendency to look at the eyes and to establish eye contact. There is neither reason to assume a compensatory behavior, producing more eye-directed gaze when the interlocutor tends to avert gaze, nor to assume a stimulating effect that would diminish or augment eye-directed gaze in accordance with the interlocutor.

Comparing the behavior of our target subject to the other subjects, she has a relatively distinct preference for the eyes as gaze target. It would be interesting to know how the use of a target subject with less distinct preference of the eyes would alter the gaze behavior of the interlocutors.

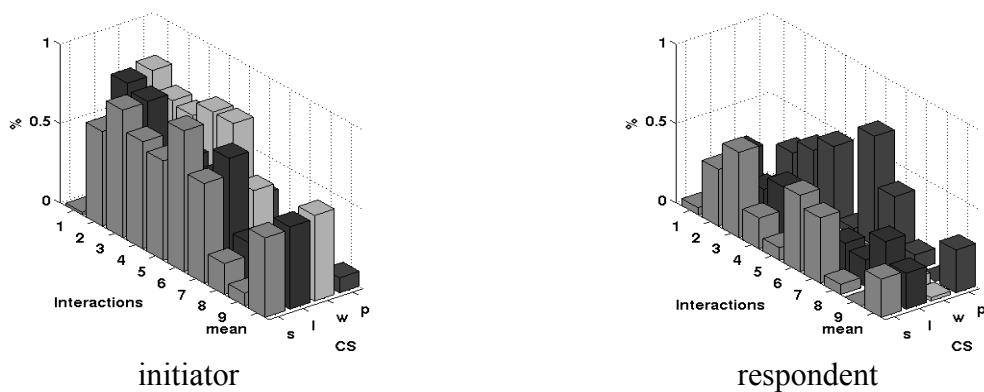


Figure 64: Percentage of mutual gaze. The durations of intervals during which both subjects direct their gaze to the others eyes are added up separated for CS and role and divided by the total duration of the respective CS. The diagram on the left shows the results for the CS speaking, listening, waiting and pre-phonation with the target subject as initiator. The diagram to the right shows the corresponding CS with the target subject as respondent.

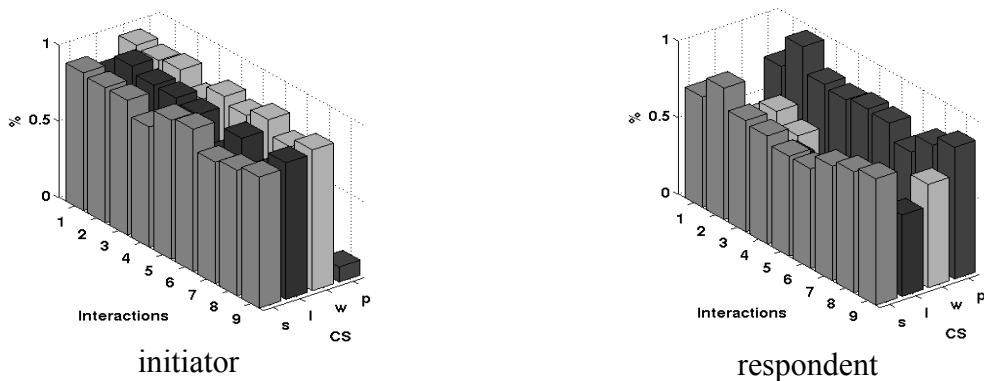


Figure 65: Percentage of the target subject's gaze directed towards the eyes of the interlocutor separated for CS and role.

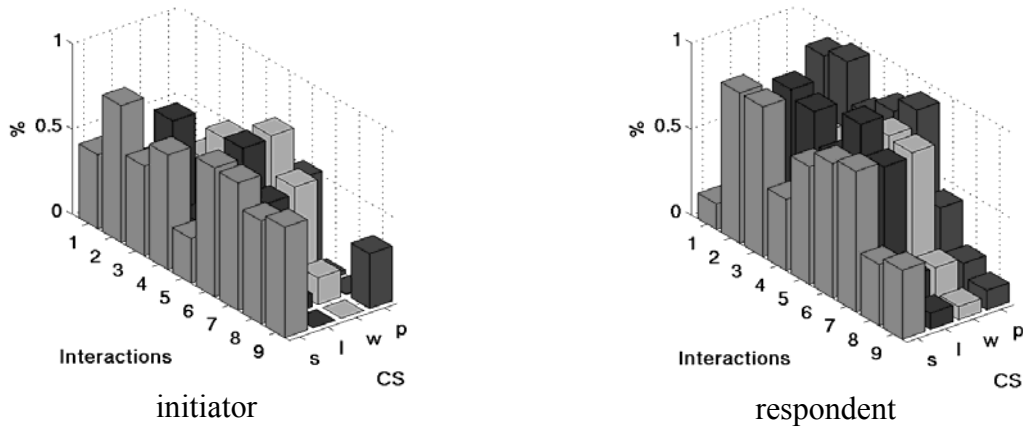


Figure 66: Percentage of the interlocutors' gaze directed towards the eyes of the target subject. The considered intervals and their separation for CS and role corresponds to the segmentation and labeling of the target subject's data.

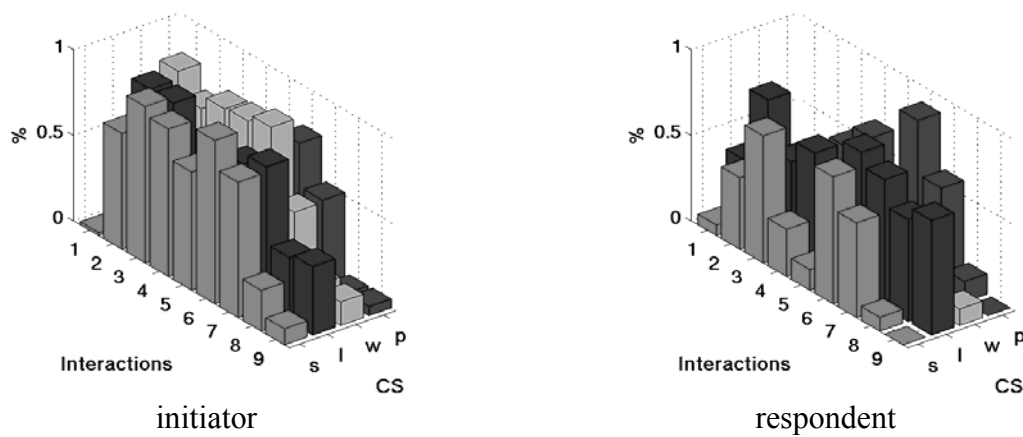


Figure 67: Percentage of mutual gaze relative to amount of eye-directed gaze of the target subject, separated for role. The amount of mutual gaze observed during the CS *speaking, listening, waiting and pre-phonation* of the target subject is divided by the amount of eye-directed gaze of the target subject observed during these intervals.

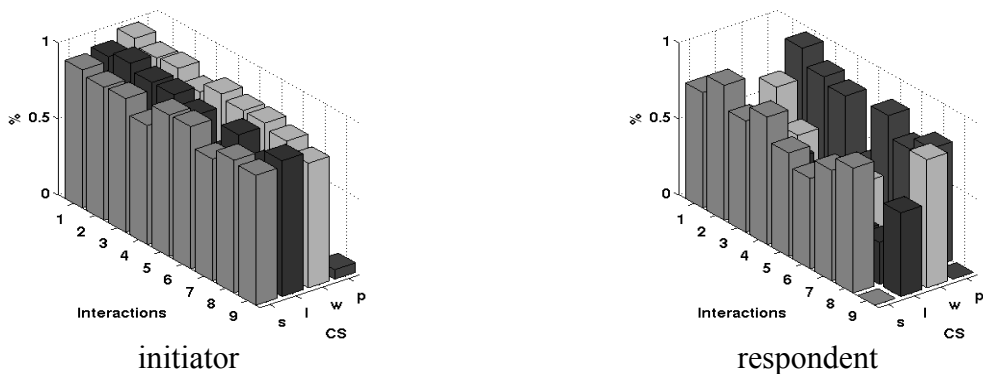


Figure 68: Percentage of mutual gaze relative to amount of eye-directed gaze of the interlocutors, separated for role. The amount of mutual gaze observed during the CS *speaking, listening, waiting and pre-phonation* of the target subject is divided by the amount of eye-directed gaze of the interlocutors observed during the respectively inspected intervals.

3.3.3 Comparison of gaze behavior in live and faked interaction

As last part of every experiment, we presented a prerecorded video to the subjects to put them into a faked interaction with a video recording of our target subject (see also section 3.1.1.3). As an explanation of the inconsistencies that may appear during the exchange of audio information during the faked interaction, the subjects were informed that during this part of the experiment, the audio feedback given to the target subject was interrupted on purpose. They were not informed that they interacted with a recording.

In the video, our target subject shows an entirely natural behavior, as it is taken from the interaction with another person. Discrepancy between the posture this person showed during the interaction that the rerecorded video is taken from, and the actual posture of the subject while interacting with this video, is often noticeable. It could in most instances be attributed to the absence of correct audio feedback.

In the following, we discuss the comparison of the data measured on six subjects interacting either with the real target subject or with its previously recorded appearance.

For visual examination of the differences of fixation time, that appear between live and faked interaction, we generated 3D bar plot diagrams. They show the mean repartition of fixation time over the six ROI we distinguish in the scene (see also section 3.3.1.1). Figure 69 shows the bar plot diagram computed for subject number 7. For the other subjects, diagrams representing the fixation time and the probability of fixations are in the appendix (section 5.4).

To determine if there are significant differences between the distributions of live and faked interaction, we used again a MANOVA test. Table 16 shows the p -values calculated for the respective pair wise comparisons of conditions for a same CS. During both of the two compared conditions, the subjects were in the role *respondent*.

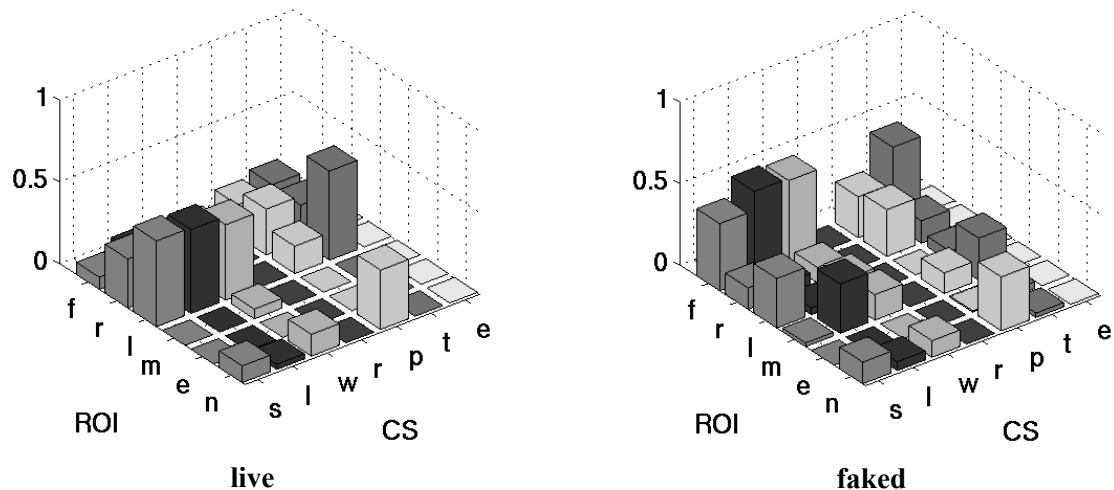


Figure 69: Mean distributions of fixation time in %/100 over ROI for the different CS computed for live interaction (left) and faked interaction (right) of subject number 7, both in the role *respondent*.

Table 16: P-values resulting from the pair wise comparison of cognitive states between live and faked interaction. Cases of significant differences are marked in boldface.

Subject	1	2	3	6	7	8
speaking	0.02	0.08	0.32	0.02	0.004	0.23
listening	0.75	0.03	0.38	0.01	5.02 e-05	0.67
waiting	0.67	0.32	0.65	0.16	2.10 e-04	0.03
pre-phonation	0.24	0.14	0.11	0.07	0.30	0.29

In some cases, there are significant differences between the distributions of the two different conditions. Only for subject 3, no significant difference occurred between the respective gaze behaviors. This subject stated that she was convinced to interact with a person and not aware that she interacted with a pre-recorded video instead. This is also true for subject 2. In contrast, subject 7, who shows very different behavior between the two conditions, was aware of this fact from the beginning due to an initial malfunctioning when starting the video playback. This may have led her to take the stimulus not serious as a person and to change her behavior accordingly.

According to our results when investigating the mutual influence between the gaze of the interacting subjects (see section 3.3.2), there is mainly endogenous control of gaze behavior, with the CS and role as factors with strong impact. No influence of exogenous factors could be found, such as for instance detailed saccadic movements of the interlocutor. We thus expect the loss of low-level coupling between gaze and video content to have minor impact, as long as the interlocutor is convinced to be able to rely on mutual attention, at least in the given scenario. We hypothesize, that the degree to which gaze behavior is different between the two conditions, is related to the degree to which a subject is aware or in doubt that the interaction is faked or otherwise impaired. However, our data is probably not sufficient to confirm this hypothesis. Only about ten samples are available per CS, as ten sentences have been exchanged. A repetition of the experiment in a similar form but with more sentences should be appropriate for this purpose.

3.4 MODELING

The statistical analysis demonstrates the influence of cognitive state and role on the gaze behavior of our target subject. These factors have an effect on fixation time, fixations length, and probability of occurrence of a fixation towards the different ROI as well on blinking rate. Based on these data we have developed a first model for mutual attention inspired by the generation process proposed by Lee *et al.* (2002) to control the gaze of our talking head. Lee *et al.* model directly the distribution of angular directions, velocities and amplitudes of saccadic eye movements (see Figure 70) according to the cognitive state of the speaker. These saccadic movements are in fact not explicitly directed to any particular ROI in the field of vision of the ECA since this information was not available in the training data. But ROI are certainly hidden in these data.

3.4.1 The HMM technique

Our model distinguishes between these two components, ROI and fixations, thanks to Hidden Markov Modeling (HMM). The essence of the HMM process is to construct a model that explains the occurrence of observations in a time sequence (see Rabiner (1989) for a tutorial). Most applications confront a set of HMMs to determine which one has most likely produced a series of unknown observation sequences. In our case, the HMMs are used for generation.

In an HMM, there are a finite number of states, each of which is associated with a (generally multidimensional) transition probability distribution. The HMM is always in one of these states. At each clock time, the system enters a state based on a transition probability depending on the previous state. After this transition is made, an output observable symbol is generated based on an observation probability distribution, depending on the current state. It is only the output symbols, not the states that are visible to an external observer and therefore states are “hidden” to the outside; hence the name Hidden Markov Model. The elements of an HMM are:

- A set of N states, $s = \{s_1, s_2, \dots, s_N\}$ with the state q at time t denoted by $q_t \in s$.
- The initial state probability distribution, $\Pi = \{\pi_i\}$, where $\pi_i = P[q_1 = s_i]$, $1 \leq i \leq N$
- The state transition probability matrix, $A = \{a_{ij}\}$, where $a_{ij} = P[q_t = s_j | q_{t-1} = s_i]$, $1 \leq i, j \leq N$
- The observation symbol probability for the observation, $B = \{b_j(O_t)\}$, where $b_j(O_t)$ is the probability of observation O_t at time t given that the state is $q_t = s_j$, $b_j(O_t) = P[O_t | q_t = s_j]$,

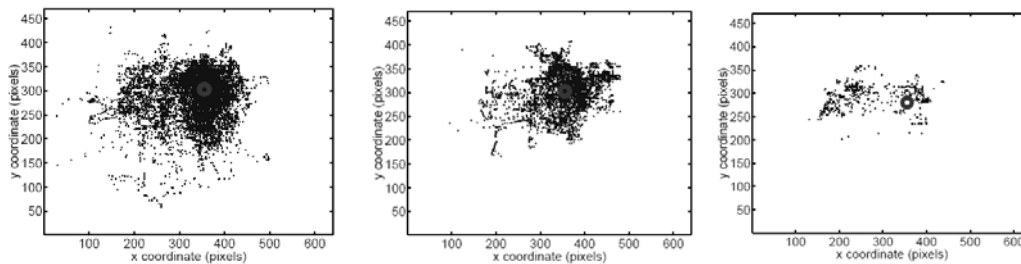


Figure 70 : The spatial distributions of saccadic eye movements in talking, listening and thinking modes (from Lee *et al.* (2002)). The circle is supposed to be centred on the interlocutor.

3.4.2 Generating fixations

One HMM takes in charge the gaze pattern of a given combination of cognitive state and role. When changing cognitive state and role, the gaze generator enrolls the appropriate HMM. In each HMM, states are simply the possible ROI and observations are fixations. The characteristics of the fixations that we retain are simply their durations. Since the distributions of the durations of the fixations depend on the ROI (e.g. fixations directed towards the mouth are longer than eye-directed fixations), the observation symbol probability B depend effectively on the state.

The initial state probability distribution reflects the fact that the target of the first fixation depends on the cognitive state and role. For all subsequent fixations, the state transition matrix A reflects the probabilities of transition of fixations between ROI. Note that the diagonal of matrix A is generally not equal to zero and correspond to what is called refixations i.e. produced by successive saccades that are directed to the same ROI. In Figure 73 a matrix of gray scale colors for each combination of CS and role represents these probabilities. Dark gray scale colors correspond to high probabilities. The colons and lines represent respectively the current and subsequent possible fixation targets. The number ‘ n ’ of accumulated fixations to a target, denoted on bottom of each colon, indicates the number of items on which this estimation is based and thus gives the reliability of the probabilities displayed in the colons of the matrix. Fixations to the face for instance are very rare. The transition probabilities from *face* to other ROI are thus not very reliable since they are calculated from only few observations.

3.4.3 Generating blinks

The initial state of each HMM emits a first fixation towards one ROI according to the initial state probability distribution. This initial state has another particular role: it also parameterizes blinking rate according to two distributions that also have been estimated using gaze data collected in the mediated face-to-face experiment. The first distribution concerns the delay of the first blink with reference to the beginning of the cognitive state: a sharp distribution around a given short delay will force a blink to occur systematically (as observed for example for pre-phonation), a large delay will inhibit blinking whereas a flat distribution will produce a random occurrence of the first blink throughout the cognitive state. The second distribution directly concerns the blinking rate: inhibition of blinking throughout the cognitive state can be simulated by a sharp distribution around zero frequency – as found in listening when being respondent - while most other HMM are characterized by Gaussians centred at around 0.3Hz in our data.

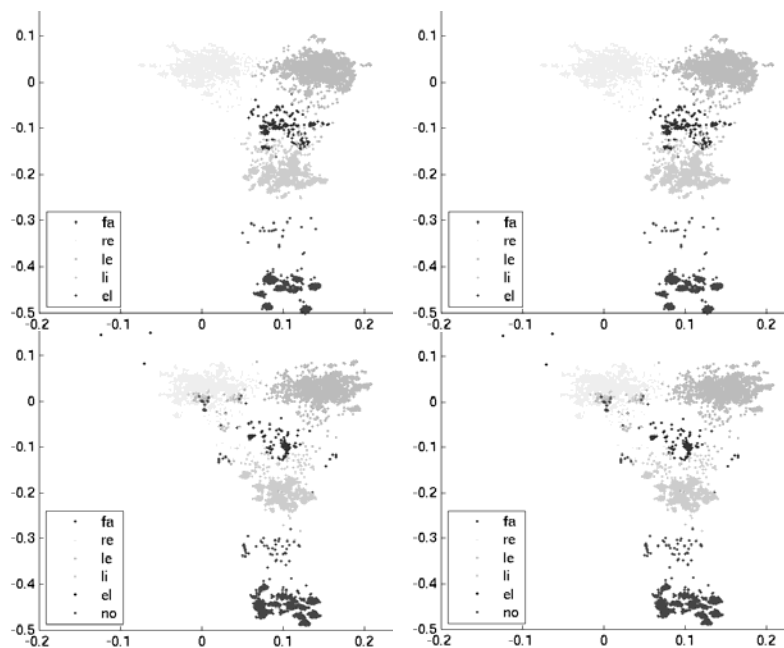
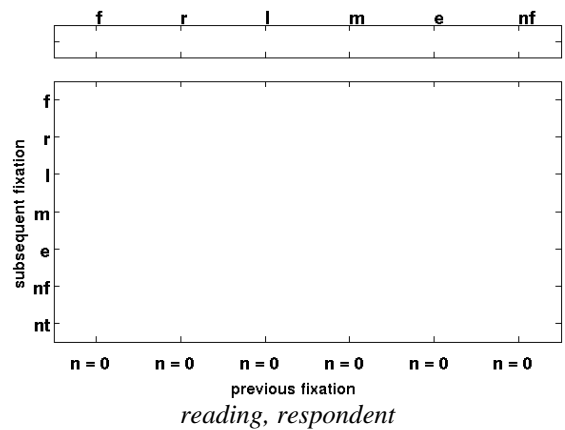
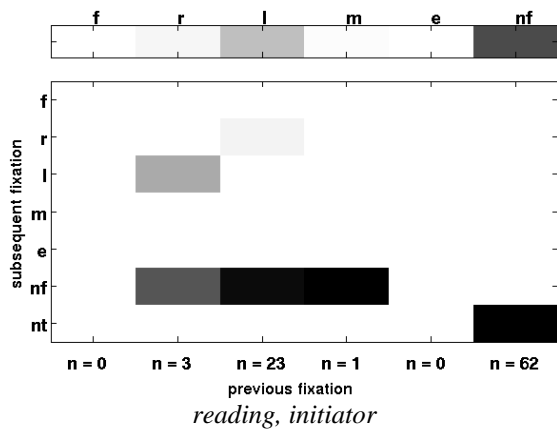
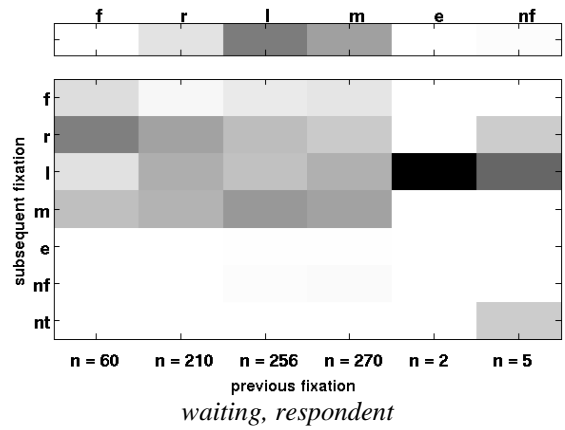
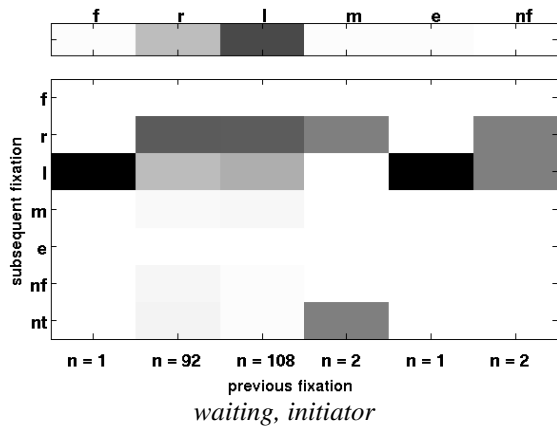
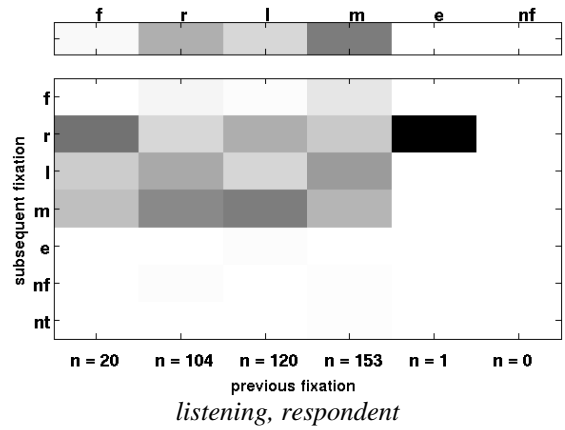
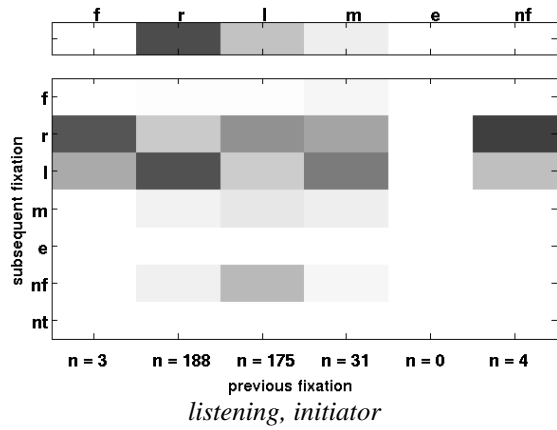
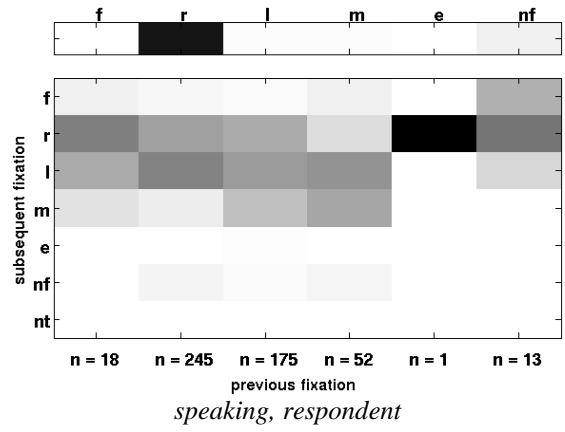
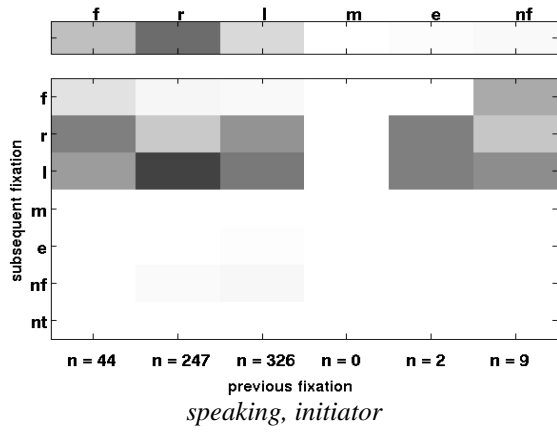


Figure 71: Comparing original gaze patterns (top) with patterns generated using the statistical model driven by the sequence of cognitive states (bottom). Left: fixations labeled with ROI; right: fixations labeled with cognitive state.



Figure 72 : Our virtual talking face is driven by 12 facial degrees-of-freedom (see Gérard Bailly *et al.* (2006a)). The eyes and eyelids movements are controlled by 5 degrees-of-freedom that captures the correlations between gaze and eyelids deformations (Gérard Bailly, Elisei, Raidt, Casari & Picot (2006b)).



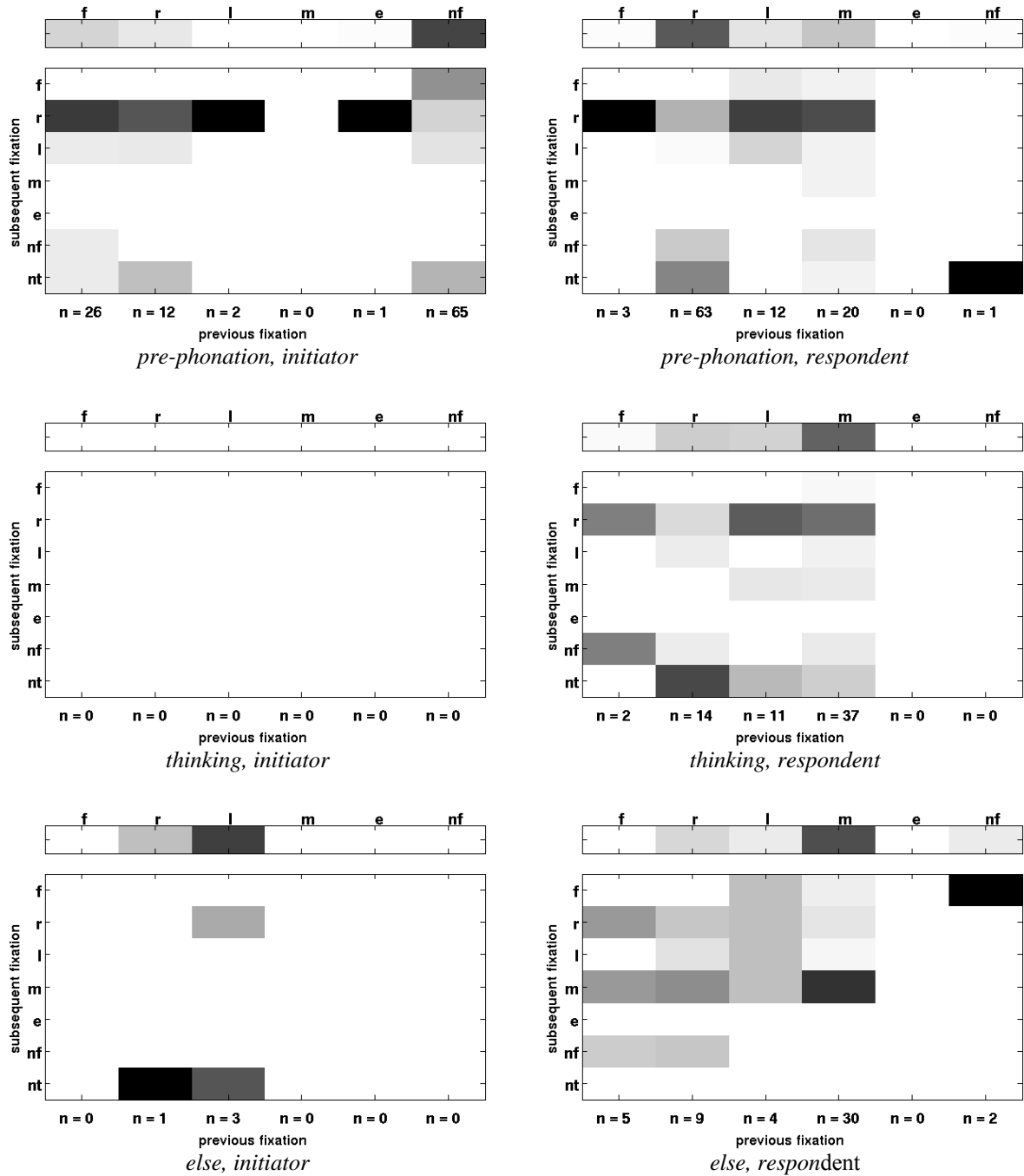


Figure 73 : Entrance and transition probabilities. For each combination of CS and role a table of entrance and transition probabilities is calculated. The upper bar of each figure displays the probability at which the first fixation of an instance of a CS is directed to a given ROI. In the array below, the transition probabilities between ROI are displayed. The columns represent the previous, the rows the subsequent fixation target. At the cross points, the probability of transition from one to another target is indicated, coded as a gray scale level, whereas the darker the color, the higher is the probability. To indicate the reliability of these values, below the diagram, the number *n* of instances from which these values are calculated is given. A small number of samples results in less reliable probabilities. The diagram covers the fixation targets *face*, *right*, *eye*, *left eye*, *mouth*, *else* as well as *no fixation* detected and *no transition*, which are indicated as abbreviations on the margins.

3.4.4 Evaluation

Until now, we have not yet evaluated the model experimentally but the distributions of fixations according to ROI and cognitive state obtained with this gaze control model are – as expected – very similar to the distributions observed during live face-to-face interactions (see Figure 71).

4 CONCLUSIONS

We performed two independent blocks of experiments, both dedicated to the investigation of the functions of gaze in dyadic conversational interaction. The first block of experiments used an already existing talking head, capable of articulatory facial movements and orientation of the head and the eyes. The aim was to verify the deictic capacity of gaze gestures as a joint movement of head and eyes. As a variation, these gestures were accompanied by speech. Whereas the gesture mainly transmits the information about the location, the concurrent speech can serve to specify the timing when exactly to evaluate this information.

The second block of experiments was designed to measure human behavior in order to acquire new knowledge about the use of gaze in dialogue. This knowledge should be exploitable to enhance the communicative capacity of the talking head with a coherent model of gaze animation. In this perspective the here presented work is only a first step. The investigated interactions represent a very restrained and hence a very basic form of dialogue in order to reduce the complexity of interaction and to obtain sufficient data of distinguishable categories. The experiments produce significant results and claim for further research on other aspects of communicative interaction in order to develop more generic models for gaze animation.

4.1 GAZE DIRECTION AND ATTENTION

In the experiment testing the deictic capacity of our talking head, in a virtual card game different conditions of assistance by a talking head were realized. From a choice of eight possible target cards, the card matching the play card had to be selected, and the play card placed on it with the help of the computer mouse.

From the results, we conclude that the talking head is able to assist subjects in an on-screen search and retrieval task. The deictic gestures provide information appropriate to reduce processing time and the effort of the search. Even if the reduced number of objects inspected during the search, do not necessarily result in diminished processing time, we still consider it as an advantage in the form of reduced cognitive load.

A very interesting point is the improvement of performance observed when the deictic gestures were enhanced by speech commands. This is probably due to the additional specification of the timing, when exactly to interpret the information of location provided by the gesture. The fact that the utterance addresses with audition another perceptual sense than the visible gestures probably contributes further to intensifies the attraction of attention towards the talking head and the information it conveys. According to the results from the questionnaire, this enhancement of multimodality by speech increases the naturalness with which the talking head is perceived.

Overall, the objective measures of performance are coherent with the subjective evaluation given in the questionnaire concerning the improvement of the talking head when gestures are accompanied by speech. They are however not necessarily coherent with respect to the individual subjects. Subjects may give better ratings in the questionnaire, but perform worse according to objective measures or the other way round. Even the objective measures may differ in the evidence they give. A reduced search path for instance does not automatically produce reduced reaction time. This confirms the subjective impression of the experimenter that the subjects applied very different strategies to accomplish the given task.

Resuming our experiments on gaze direction and attention, there arise several perspectives for further improvement of the animation of the talking head as well as for improvements and modifications of the experimental scenario.

In a condition, where subjects could only rely on the deictic gestures, without the option to identify the correct target card from visual inspection, several confusions between neighboring targets occurred. This gave place to an improvement of the modeling of shape of eyes and eyelids in order to provide correct visual stimuli. The eyelids experience deformations due to the movements of the eyes both in vertical and horizontal direction. The realization of these deformations for the animation of the talking head, should render the appearance of the eyes much more realistic and provide more precise information on the exact orientation of the eyes. Unfortunately, because of lack of time, the implementation of this improved model has not been tested yet.

According to Fitts's law, distance and size of object are known to influence performance in pointing and selection (Surakka *et al.* (2003)). This should be considered for the improvement of our experimental scenario, especially when controlling the precision of interpretation of gaze direction from the enhanced model of eyelids. The diminishing of object size and the extension of their number should allow for results that are more refined.

The multimodal deictic gesture could furthermore profit from a detailed measurement of the coordination of such gestures observed in human interaction. The actual rendering and timing of the gestures has not been backed up by experimental data. This would also afford an improvement of the finite state automaton to enable a more exact coordination of actions in the measure that this is possible with the available operating system.

4.2 MEDIATED FACE-TO-FACE INTERACTION

To investigate gaze behavior in close dyadic interaction, we developed a setup for mediated face-to-face interaction and conducted experiments with a scenario limiting interaction to a basic conversation of uttering and repeating of imposed sentences. The outcome of the experiment confirms the scenario and setup as appropriate means to investigate the gaze behavior of subjects with the possibility to relate it to the ongoing interaction. The cognitive states and conversational roles we distinguish proved to have strong influence on gaze and blink behavior. The analysis of the acquired data allowed for a basic model of gaze animation for an embodied conversational agent in dyadic conversation.

The repartition of fixations clearly confirms the findings reported in literature, with the eyes and the mouth as prominent targets (Vatikiotis-Bateson *et al.* (1998), Lansing & McConkie (1999)). Whereas the individual behavior of the different subjects varies in the repartition of gaze over these targets, our target subject showed a rather consistent and balanced repartition of gaze over the eyes and the mouth during the nine analyzed interactions in which she participated. In certain cases, there is a very obvious influence of cognitive state and role on these patterns. Our target subject shows a tendency to look at the mouth while listening that becomes very strong in the role respondent, where the listener receives new information. This confirms the argumentation in Lansing & McConkie (1999) that attention to word content produces more mouth directed fixations. However, even in the cases, where the mouth is an important source of information, the eyes are still important targets and are fixated at high probability. This reflects the importance of the social aspect of gaze behavior in which eye-directed gaze plays an outstanding role.

The search for relations between the gaze behaviors of the interacting subjects in the sense of mutual influence did not produce exploitable data. There was for instance no relation between the gaze behavior and the occurrence of eye contact. We expected either a search for eye contact or its avoidance. That no such relation could be found might be due to the scenario and the use of imposed sentences, restricting the content of communication as well as its structure.

The observations reported by Kendon (1967) cannot be confirmed with our data. He observed patterns of averted gaze and face-directed gaze in the context of turn taking. Averted gaze is of no importance in our data, as close to 100% of the fixations are directed to regions on the face. The fact that the scenario imposes the turn taking in our experiments may reduce the bargaining for the turn and weaken the peculiarities of behavior observed in this context. The conditions of the two experiments are not comparable and we do therefore not consider the different observations as contradictory.

Compared to the method applied by Kendon, our level of resolution of gaze direction is more detailed, as well as the structuring of conversation with cognitive states. A behavior comparable to his observations in the context of turn taking might be the frequent blink during the CS *pre-phonation* and the strong tendency to fixate the right eye during this CS.

In general, the analysis of blink evidences also a relation between frequency of occurrence and cognitive state and role. The observations favor the hypothesis that increased attention leads to reduced frequency of blink. This may either be a social signal of attention, or due to the necessity not to impede view and visual perception while receiving information. Most probably, it is a combination of both. Furthermore, there is a remarkable tendency to blink when preparing to speak, especially after the CS *reading* in the role *respondent*. As it occurs usually during the upward movement of the head, from the gaze towards the written sentences on the table, up at the interlocutor, this may be a protective reflex (Evinger *et al.* (1994)). Most probably it is however a signal in the context of turn taking, similar to the aversion of gaze at the beginning of a turn, reported by Kendon (1967). He considered such aversion of gaze as a strategy to insist on the turn, preventing competing signals from the interlocutor.

The results validate also our experimental setup as such, which is the first one to our knowledge that enables detailed monitoring of both subjects in dyadic interaction. The results obtained with a rather restricted scenario of interaction, encourage the extension of the experiments onto further aspects of conversation. The scenario should develop stepwise towards unrestricted interaction, permitting to further detail the coarse basic relations we discovered. The use of ordinary sentences instead of SUS, variation of the task or variation of the relation of sex and social status may be such options. Furthermore, it should be possible to inspect additional cognitive states or attitudes (doubt, surprise, etc.) not addressed in the current scenario.

Independent of the progress in measurement of gaze behavior, the derived models need verification. The model we propose should be tested using an ECA in the given scenario to see how it influences the behavior of the subjects compared to the interactions with our target subject. The hypothesis that frequency of blink declines with increasing attention, could be verified in a similar scenario, inverting this relation. If blink increases when attention is expected to be high, this may irritate the interlocutor and produce visible reactions.

The model of mutual gaze that emerges from our monitoring of mediated human-human interactions should be confronted to human-agent face-to-face interaction. Task-oriented dialogs such as involved in games offer an interesting framework for measuring the impact of increased mutual attention on performance. The card game used in the experiment on multimodal deixis can be extended to include verbal negotiations between the ECA and the subjects so that to combine gaze skills for mutual attention and deixis in a unique collaborative task.

A long term goal is of course the integration of the gaze model for face-to-face interaction, the use of coordinated head and eye movements as deictic gestures as well as the gaze model for scene perception (Picot *et al.* (2007)) as a single complete model for the animation of gaze of our talking head. A comprehensive modeling of dialogue would complete our talking head to become an operational embodied conversational agent.

5 APPENDIX

5.1 QUESTIONNAIRE PRESENTED ON SCREEN AFTER THE CARD GAME

Experiment I

- Lorsqu'aucun clone n'était affiché, avez-vous eu l'impression de pouvoir répondre rapidement ?
- Jugez-vous les mouvements du clone comme réalistes ?
- Lorsque le clone montrait la mauvaise carte, avez-vous eu conscience de regarder le clone ou de suivre ses indices ?
- Lorsque le clone montrait la bonne carte alors que sa valeur était visible, avez-vous eu conscience de regarder le clone pour trouver la carte ?
- Lorsque le clone indiquait la bonne carte alors que sa valeur était visible, avez-vous eu l'impression de pouvoir répondre rapidement ?
- Lorsque le clone indiquait la bonne carte alors qu'elles restaient toujours masquées, avez-vous trouvé que le clone regardait vers la bonne carte avec précision ?
- Selon vous, pour quelle tâche avez-vous été plus rapide ?
- Si vous aviez le choix pour faire cette tâche, quelle condition choisiriez-vous ?
- Quel est votre œil directeur (D/G) ?

Experiment II

- Jugez-vous la manière du clone de donner des indices comme réaliste ?
- Lorsque le clone désignait la mauvaise carte, avez-vous regardé le clone et suivi ses indices ?
- Lorsque le clone désignait la mauvaise carte, cela a-t-il dérangé votre recherche ?
- Lorsque le clone désignait la bonne carte, avez-vous regardé le clone pour trouver la carte ?
- Lorsque le clone désignait la bonne carte, avez-vous eu l'impression de pouvoir répondre plus rapidement ?
- Selon vous, pour quelle tâche avez-vous été plus rapide ?
- Si vous aviez le choix de l'interface, quelle condition choisiriez-vous ?
- Quel est votre œil directeur (D/G) ?

5.2 SETUP CHECKLIST: MEDIATE FACE-TO-FACE INTERACTION

Préparations pour une expérience avec le setup**Face-à-Face****Sans sujet**

Oculomètres :

- Contrôler le bon branchement sur la bonne machine
! Ne pas changer les branchements de l'oculomètre quand lui ou sa machine sont allumées !
(Vérifier le branchement de l'oculomètre comme 2^{ème} écran via le boîtier)
- Assurer que les oculomètres sont réveillés (lancer un faux enregistrement)
- Si nécessaire, rebooter sous Windows pour que l'oculomètre sois reconnu par le système
- Vérifier le fonctionnement des oculomètre (lancer un faux enregistrement)
- Vérifier les settings de ClearView (voir annexe)
- Laisser enregistrement de CV en attente
- Brancher et tester l'éclairage, débrancher ensuite

Vidéo / Audio :

- Installer les cameras de manière pas croisé sur les écrans
- Vérifier le fonctionnement des cameras et les settings de la carte d'acquisition avec VirtualDub (voir annexe)
 - 25 FPS
 - Couleurs YUY2
 - Pal_B
 - Taille : 320 x 240
 - Audio : 22 kHz, 16 Bit, Stéréo (Bronze : entré AUX ; Cuivre : entré Ligne)
- Spécifier un fichier pour l'enregistrement vidéo sur Bronze avec la commande 'Set Capture File' dans VirtualDub
- Brancher les microphones
- Allumer les préamplis Sony
- Allumer l'Ampli Micro
- Allumer les enceintes et les tester (croisés)
- Tester les paramètres de la carte son (VirtualDub ou Praat ou panneau de config) (dispositif d'enregistrement doit être l'entre ligne)
- Reteindre les préamplis Sony (économiser les piles)
- Ajuster la résolution du boîtier 'Vidéo Console'

Synchronisation :

- Charger le driver « giveio » dans répertoire local `\Raidt\Face-a-Face\TET_sync`
 - nécessite éventuellement être logué comme Administrateur sur Cobalt
 - sur Cuivre c'est possible de le lancer via un terminal, voir 'Charger_giveio.txt'
- Modifier le nom du Sujet dans le fichier: `\Raidt\Face-a-Face\TET_sync\Sujet_TET_sync`
- Insérer le nom de la calibration de ClearView à la place de `<nom_calib CV>`

Debug\TET_synch.exe -tet <nom_calib CV> -screen 1280x1024+1280+0 -beep – starter

Debug\TET_synch.exe -tet <nom_calib CV> -screen 1280x1024+1280+0 -beep – starter

(attention à la résolution de l'écran)

- Brancher les émetteurs du signal de synchronisation de manière croisée et les tester avec *'Func_test_TET_Synch.bat'*
- Préparer sur une machine sous Linux *'cv_starter'* dans le répertoire *'/stordata/raidt/Face_a_Face'*
 - Exemple: *cv_starter –synch Cuivre Cobalt*
 - En cas d'un seul oculomètre ajouter sa machine comme premier paramètre

Avec sujet présent

Vidéo / Audio :

- positionner le sujet dans l'espace surveillé par l'oculomètre (moniteur de CV)
 - Ajuster la position de l'oculomètre et l'hauteur de la chaise pour que le sujet soit dans une position confortable
- faire une calibration de CV sous le nom du Sujet
- allumer et positionner l'éclairage
- allumer le boîtier vidéo
- Ajuster l'orientation des camera en branchement pas croisé pour qu'elles soient centrée entre les yeux légèrement au dessus. Eventuellement ajuster l'emplacement des cameras.
- Croiser les cameras et vérifier l'affichage des visages par rapport aux cameras
- Allumer les préamplis Sony

Synchronisation :

- Lancer l'application de synchronisation *TET_sync* sur les deux machines
- Positionner les émetteurs du signal de synchronisation en champ de vision des cameras
- Démarrer l'enregistrement avec VirtualDub sur Bronze avec **F5** (arrêter avec **ESC**)
- Démarrer *'cv-starter'*

Après l'expérience

Sauvegarde de données :

- Générer un répertoire dans *'/stordata/raidt/Face_a_Face/_exp_face_a_face/'* avec le nom *<Subject_1>_<Subject_2>* et ses sous-répertoires (use copy of Template) □
 - Déplacer les fichiers *'<Subjct_n>_TET_gz.txt'* avec les donnée de TET_sync □
 - Exporter de CV les fichiers (attention au 'Study Settings') :
 - *<Subject>EFD.txt*
 - *<Subject>GZD.txt*
 - Copier les fichiers d'enregistrement de CV □
 - *Audio.mp3*
 - *ExternalVideoRec.avi*
- du répertoire :
- C:\Program Files\Tobii\ClearView\Application Data\ClearViewData\Studies\ ... \Recordings'*
- ou bien le AVI enregistré avec CV sur 'Bronze'
- Générer avec 'Nero Wave Editor' le fichier *<Subject_1>_<Subject_2>.wav* dans le répertoire *'<Subject_1>_<Subject_2>'* de préférence à partir du fichier *'Audio.mp3'*

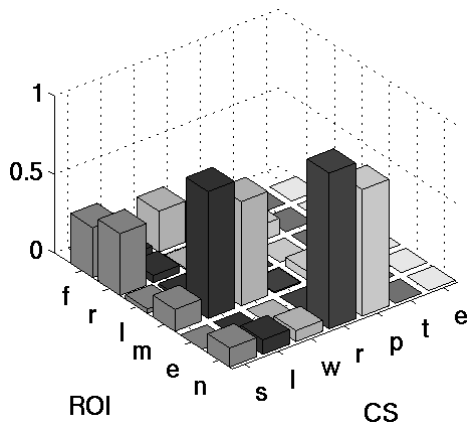
Détermination des paramètres (avec Praat et Virtual Dub)

- Avec Praat, lire le fichier *<Subject_1>_<Subject_2>.wav* comme fichier stéréo et détermine les débuts de beeps sonores.
- Avec VirtualDub, lire les fichiers AVI :
 - Sélectionner la séquence à extraire et noter les numéros d'image de début et fin
 - Extraire la séquence avec commande 'Save image sequence ... ' et sauvegarder dans répertoire 'ExtractedImages' du sujet.
 - Déterminer la période Master de chaque sujet

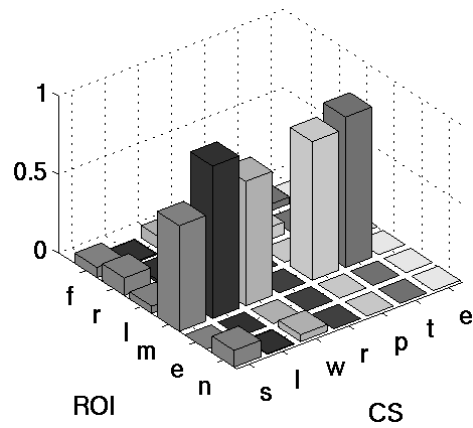
<i>Subject_1</i>	Start		Stop
Master			
Video Sync			
Video Sequence			
Audio Sync			

<i>Subject_2</i>	Start		Stop
Master			
Video Sync			
Video Sequence			
Audio Sync			

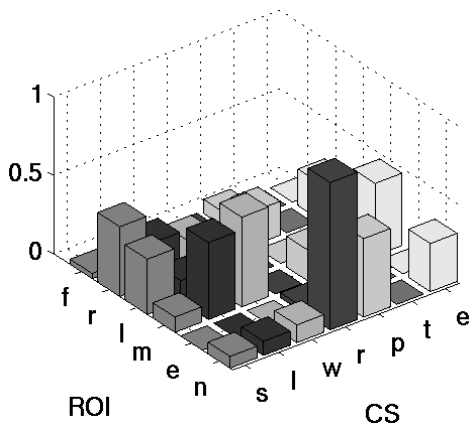
5.3 FIXATION TIME AND FIXATION PROBABILITIES OF INTERLOCUTORS OF OUR TARGET SUBJECT



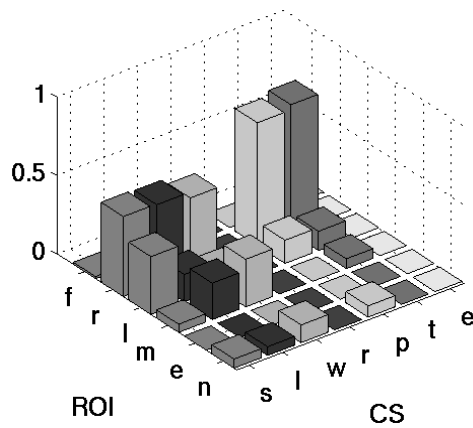
**Subject 1 – initiator;
fixation time in %/100**



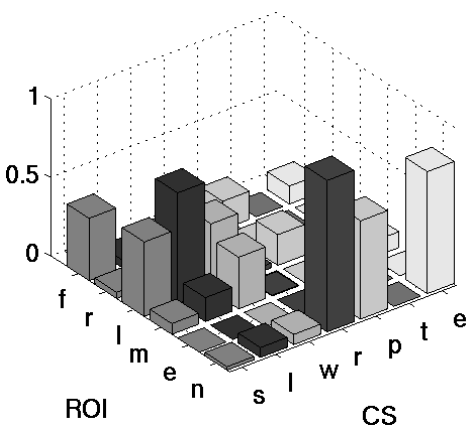
**Subject 1 – respondent;
fixation time in %/100**



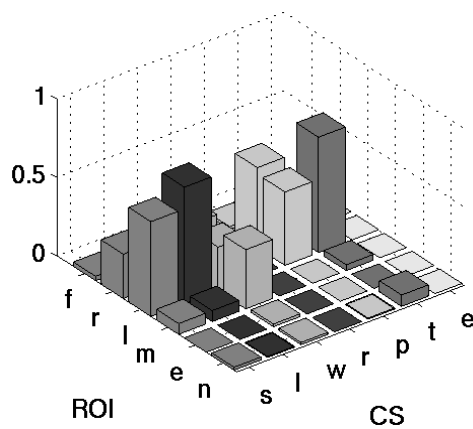
**Subject 2 – initiator;
fixation time in %/100**



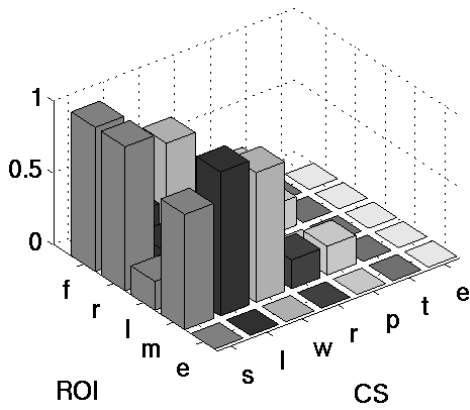
**Subject 2 – respondent;
fixation time in %/100**



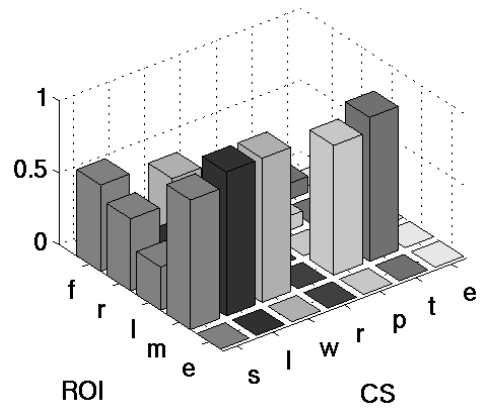
**Subject 3 – initiator;
fixation time in %/100**



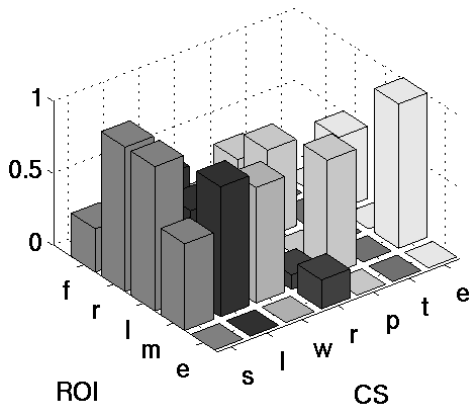
**Subject 3 – respondent;
fixation time in %/100**



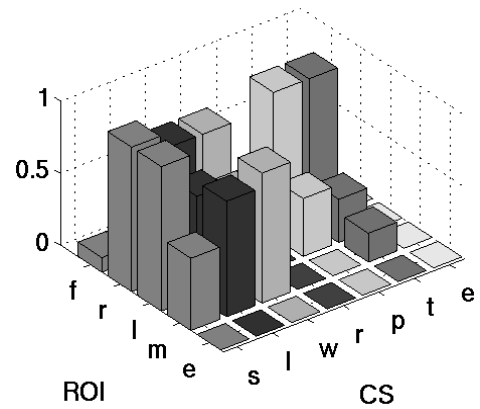
**Subject 1 – initiator;
probability of fixation in %/100**



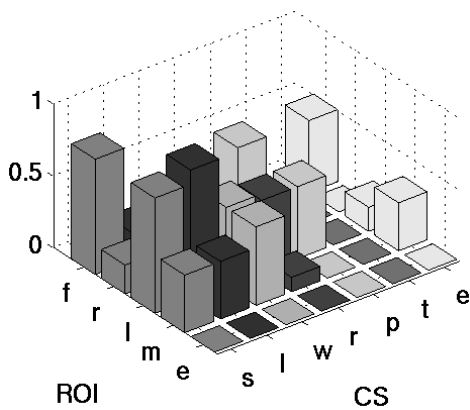
**Subject 1 – respondent;
probability of fixation in %/100**



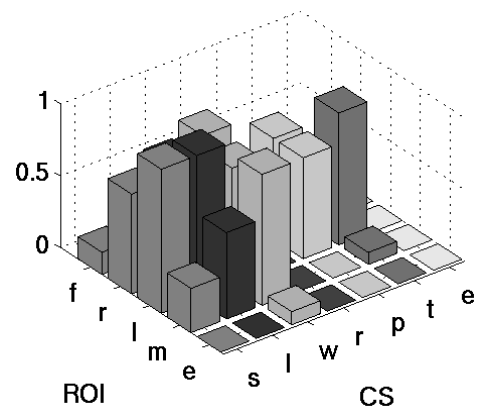
**Subject 2 – initiator;
probability of fixation in %/100**



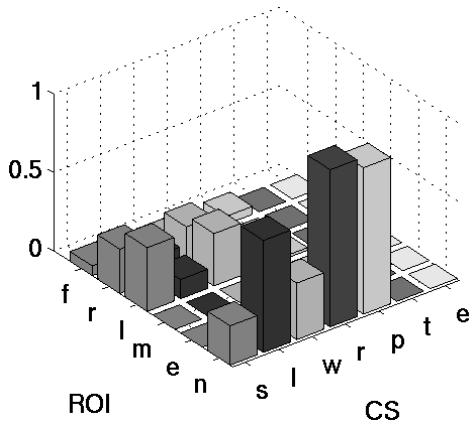
**Subject 2 – respondent;
probability of fixation in %/100**



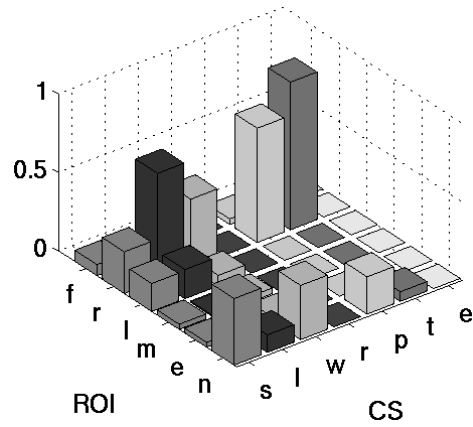
**Subject 3 – initiator;
probability of fixation in %/100**



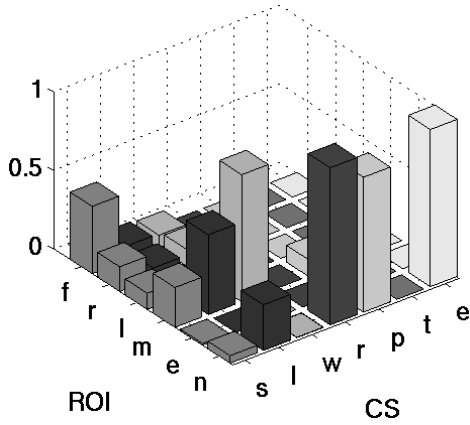
**Subject 3 – respondent;
probability of fixation in %/100**



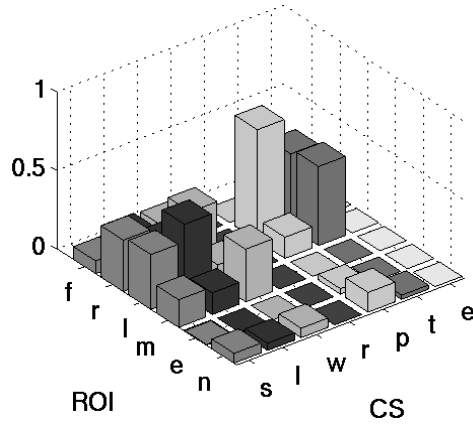
**Subject 4 – initiator;
fixation time in %/100**



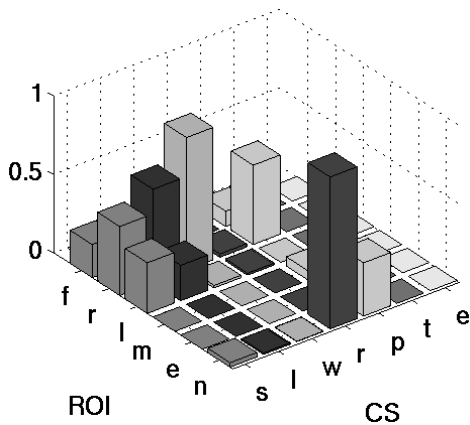
**Subject 4 – respondent;
fixation time in %/100**



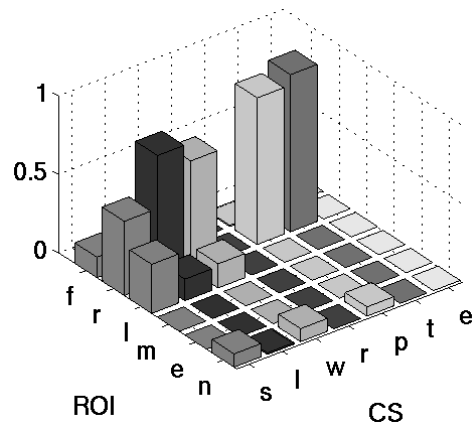
**Subject 5 – initiator;
fixation time in %/100**



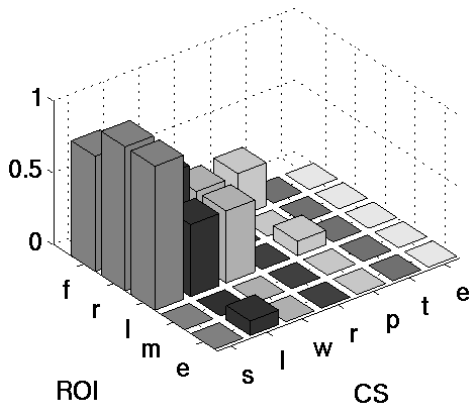
**Subject 5 – respondent;
fixation time in %/100**



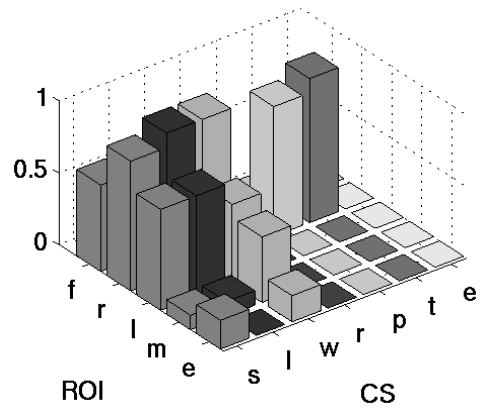
**Subject 6 – initiator;
fixation time in %/100**



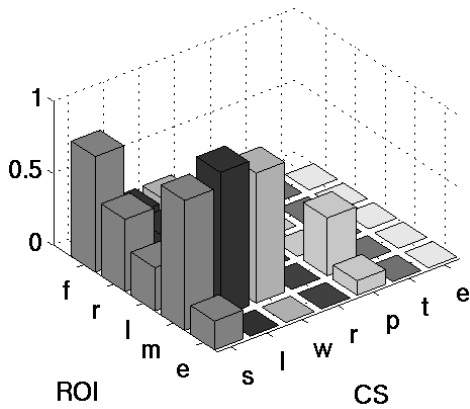
**Subject 6 – respondent;
fixation time in %/100**



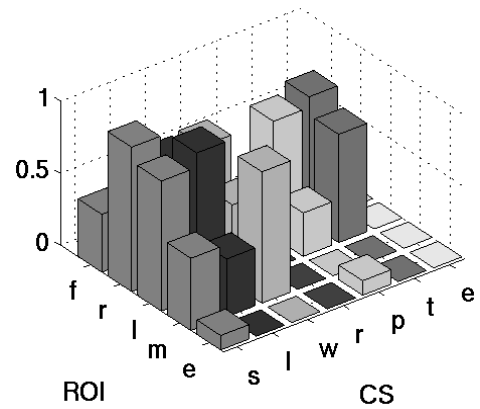
**Subject 4 – initiator;
probability of fixation in %/100**



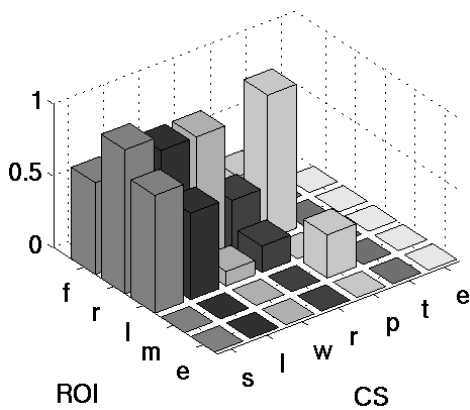
**Subject 4 – respondent;
probability of fixation in %/100**



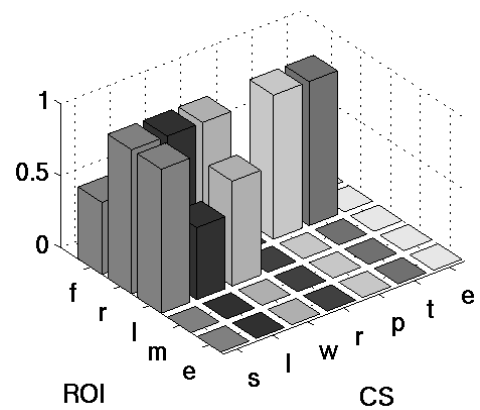
**Subject 5 – initiator;
probability of fixation in %/100**



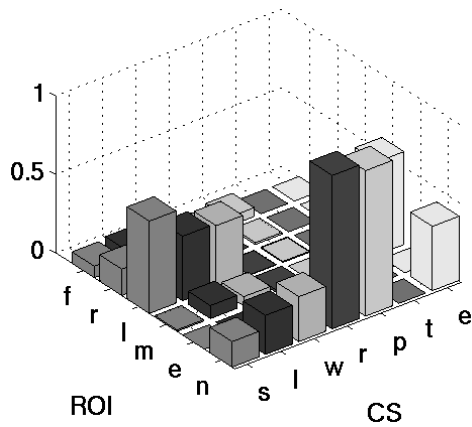
**Subject 5 – respondent;
probability of fixation in %/100**



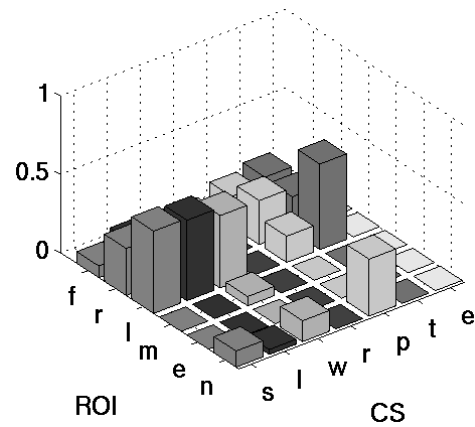
**Subject 6 – initiator;
probability of fixation in %/100**



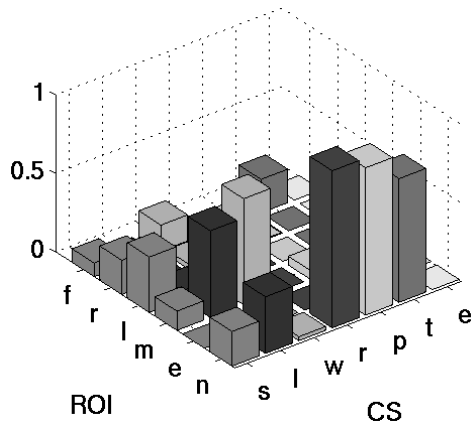
**Subject 6 – respondent;
probability of fixation in %/100**



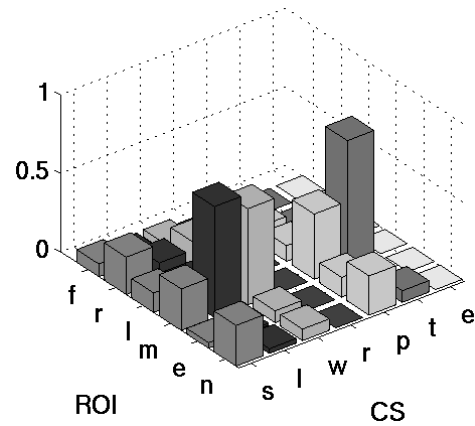
**Subject 7 – initiator;
fixation time in %/100**



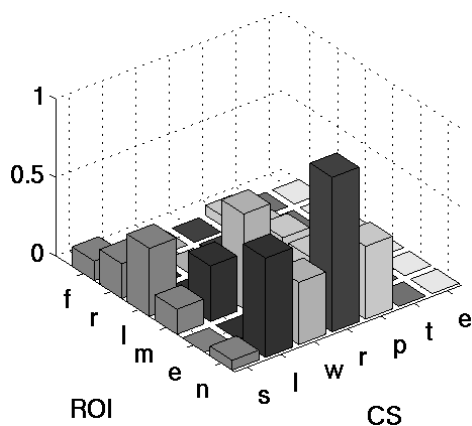
**Subject 7 – respondent;
fixation time in %/100**



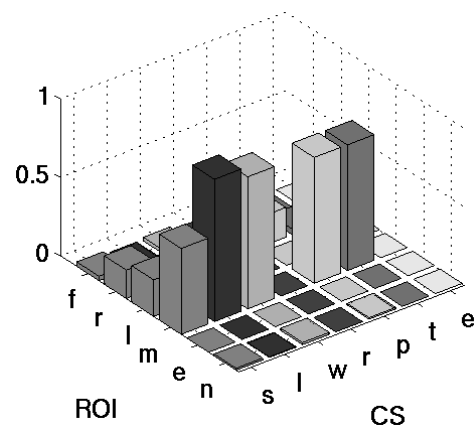
**Subject 8 – initiator;
fixation time in %/100**



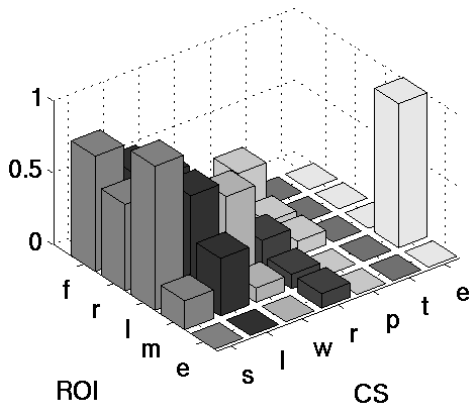
**Subject 8 – respondent;
fixation time in %/100**



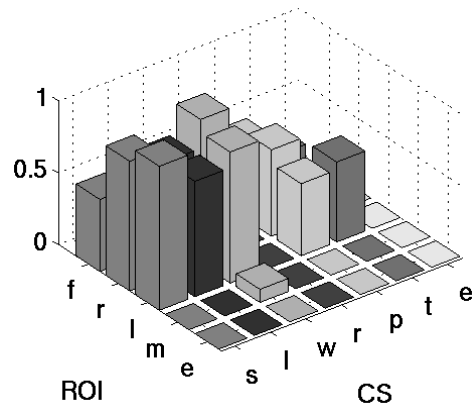
**Subject 9 – initiator;
fixation time in %/100**



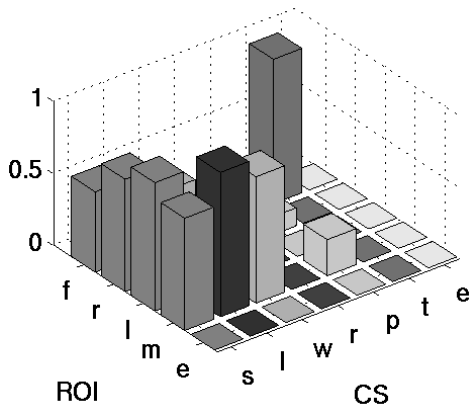
**Subject 9 – respondent;
fixation time in %/100**



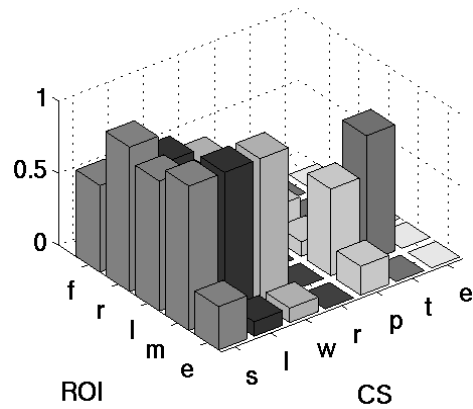
**Subject 7 – initiator;
probability of fixation in %/100**



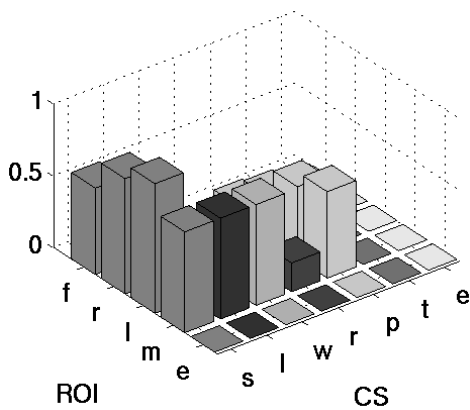
**Subject 7 – respondent;
probability of fixation in %/100**



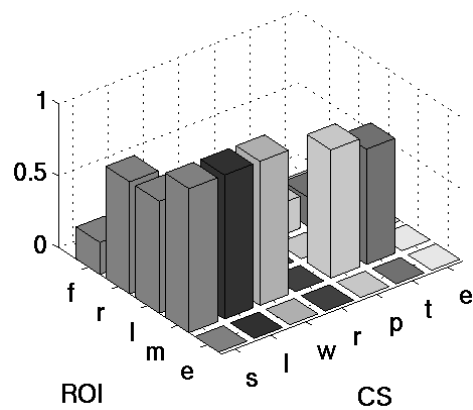
**Subject 8 – initiator;
probability of fixation in %/100**



**Subject 8 – respondent;
probability of fixation in %/100**

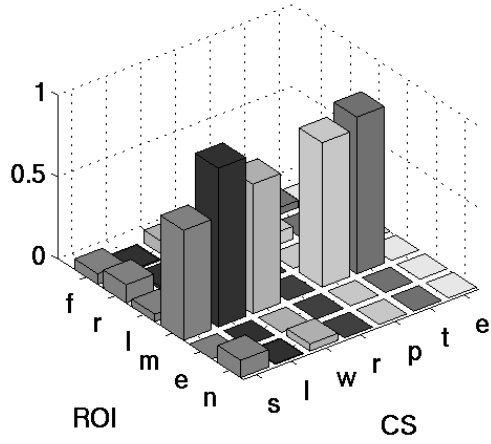


**Subject 9 – initiator;
probability of fixation in %/100**

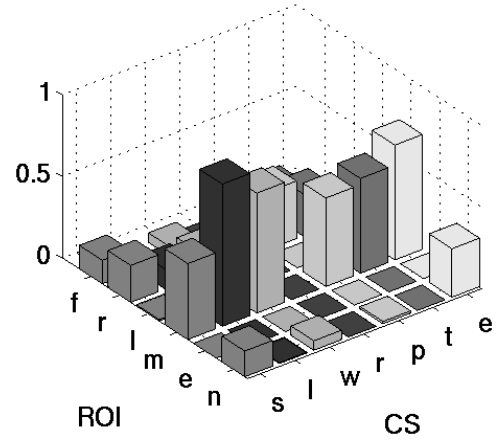


**Subject 9 – respondent;
probability of fixation in %/100**

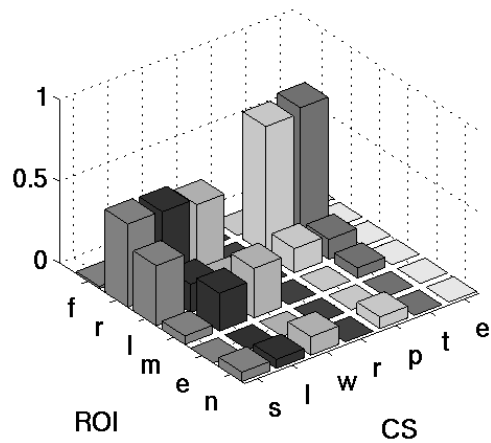
5.4 COMPARISON OF FIXATION TIME AND FIXATION PROBABILITY BETWEEN LIVE AND FAKED INTERACTION



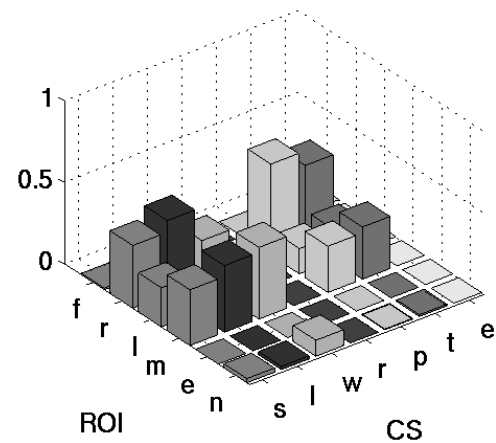
Subject 1 – respondent, live;
fixation time in %/100



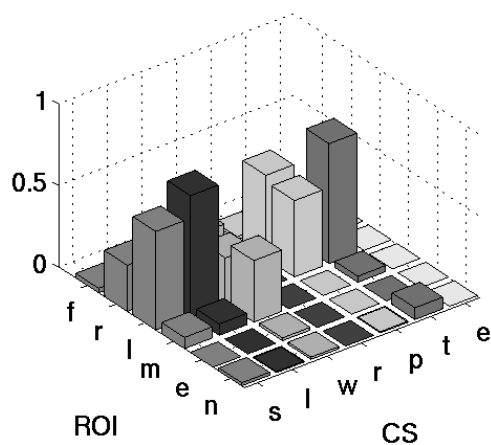
Subject 1 – respondent, faked;
fixation time in %/100



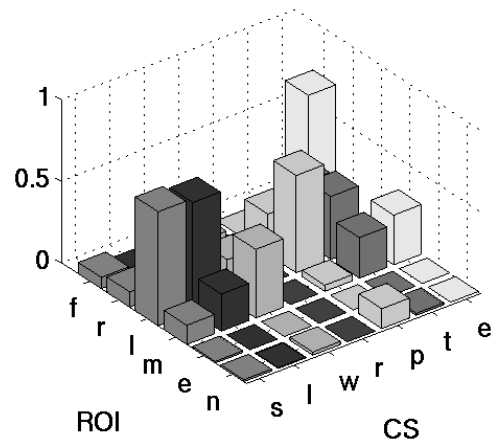
Subject 2 – respondent, live;
fixation time in %/100



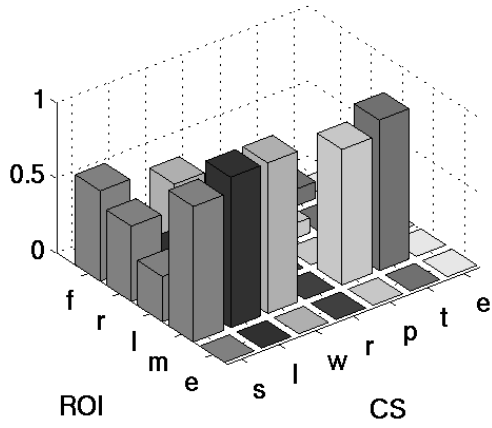
Subject 2 – respondent, faked;
fixation time in %/100



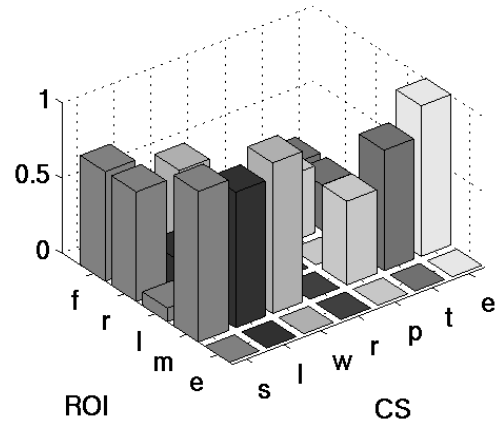
Subject 3 – respondent, live;
fixation time in %/100



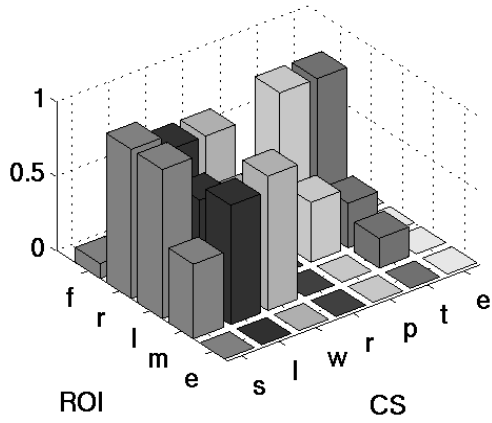
Subject 3 – respondent, faked;
fixation time in %/100



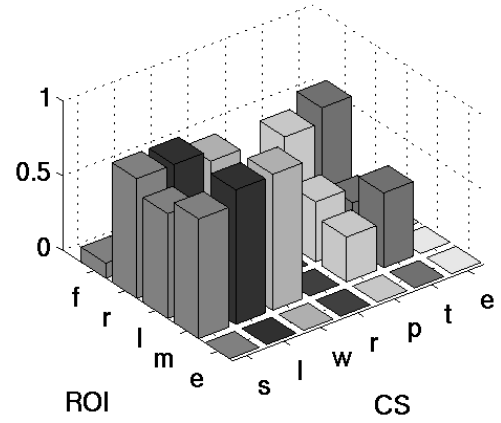
**Subject 1 – respondent, live;
probability of fixation in %/100**



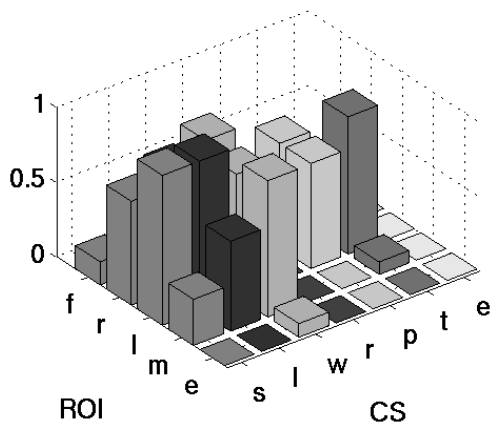
**Subject 1 – respondent, faked;
probability of fixation in %/100**



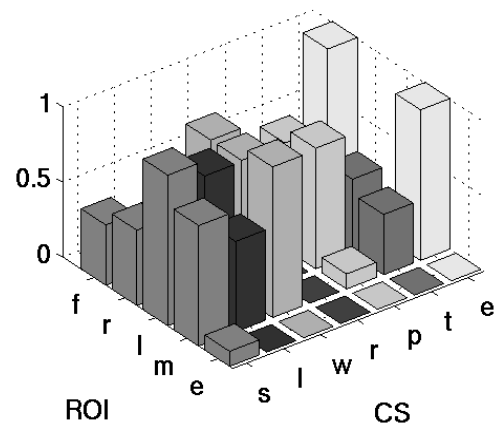
**Subject 2 – respondent, live;
probability of fixation in %/100**



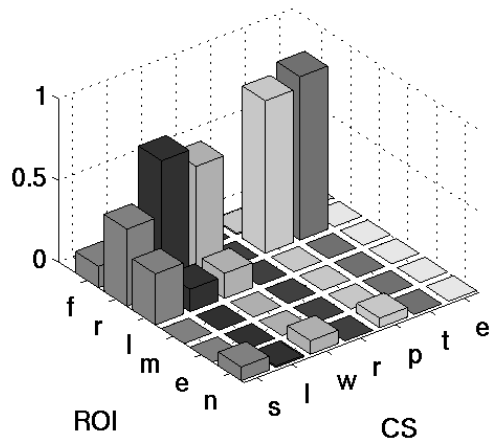
**Subject 2 – respondent, faked;
probability of fixation in %/100**



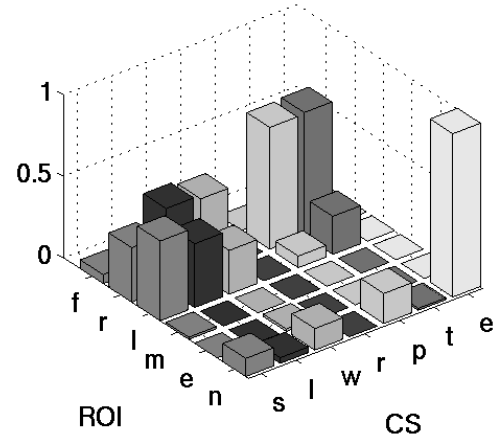
**Subject 3 – respondent, live;
probability of fixation in %/100**



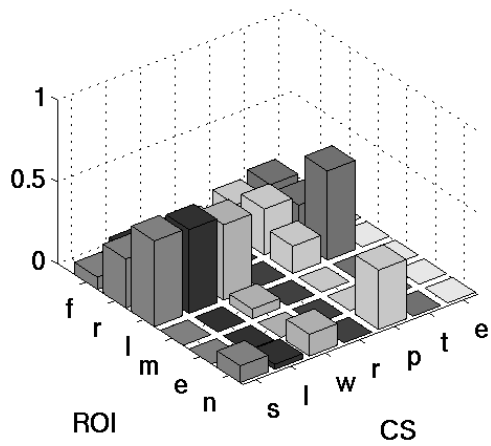
**Subject 3 – respondent, faked;
probability of fixation in %/100**



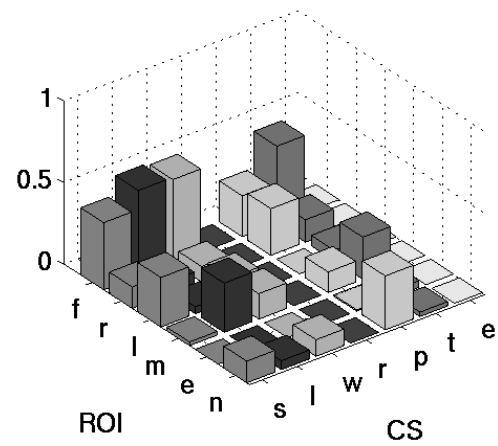
**Subject 6 – respondent, live;
fixation time in %/100**



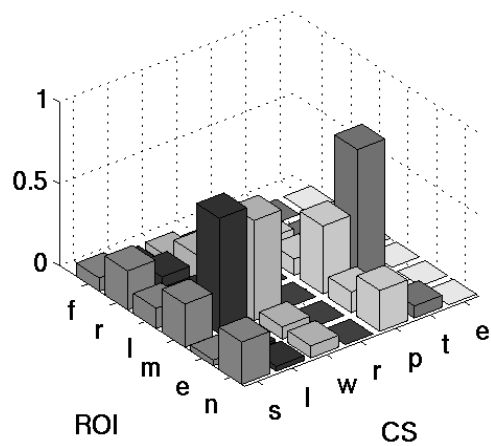
**Subject 6 – respondent, faked;
fixation time in %/100**



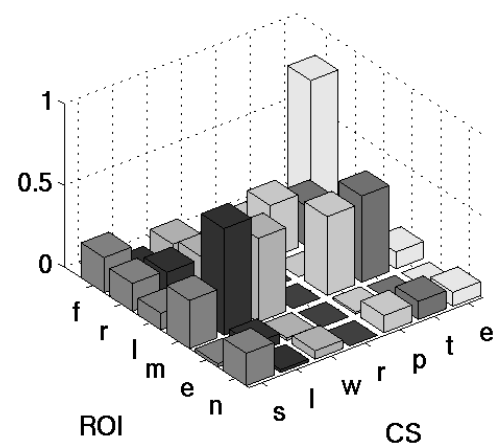
**Subject 7 – respondent, live;
fixation time in %/100**



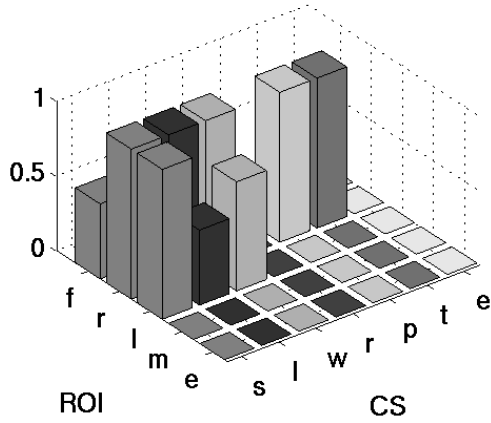
**Subject 7 – respondent, faked;
fixation time in %/100**



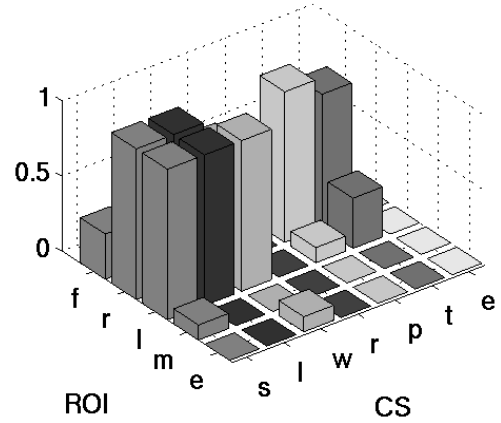
**Subject 8 – respondent, live;
fixation time in %/100**



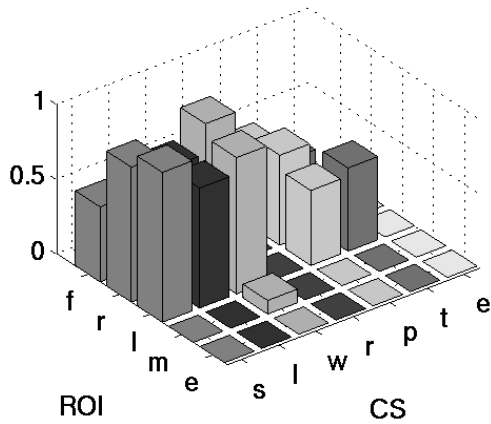
**Subject 8 – respondent, faked;
fixation time in %/100**



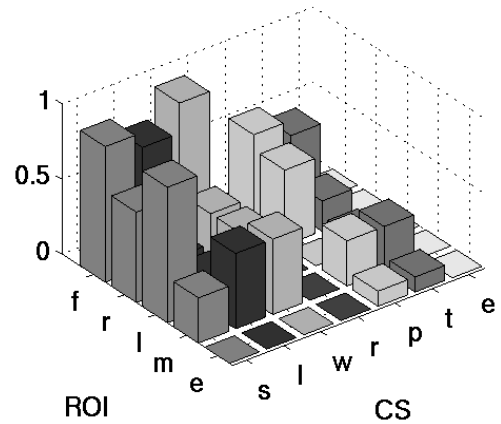
**Subject 6 – respondent, live;
probability of fixation in %/100**



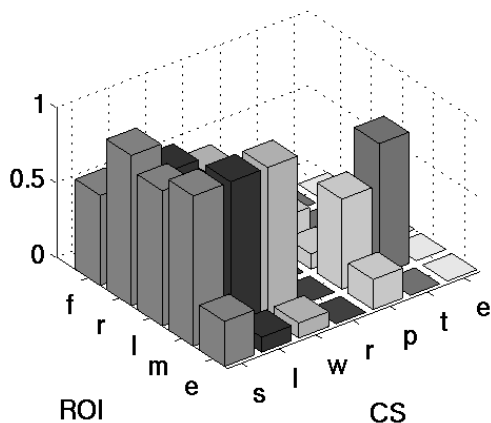
**Subject 6 – respondent, faked;
probability of fixation in %/100**



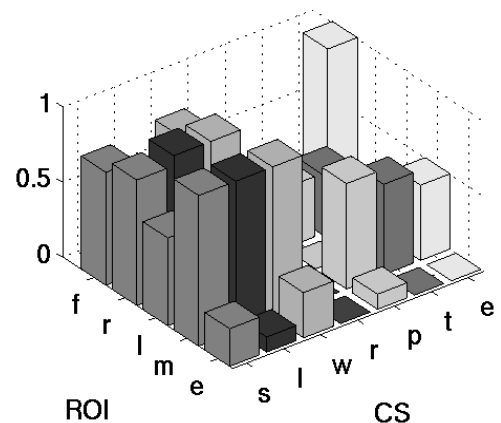
**Subject 7 – respondent, live;
probability of fixation in %/100**



**Subject 7 – respondent, faked;
probability of fixation in %/100**



**Subject 8 – respondent, live;
probability of fixation in %/100**



**Subject 8 – respondent, faked;
probability of fixation in %/100**

6 REFERENCES

- Ahrens, R. (1954). Beitrag zur Entwicklung des Physiognomie- und Mimikerkennens. *Zeitschrift für Experimentelle und Angewandte Psychologie*, **2**(3), 412-454.
- Anstis, S.M., Mayhew, J.W. & Morley, T. (1969). The perception of where a face or television "portrait" is looking. *American Journal of Psychology*, **82**, 474-489.
- Argyle, M. & Cook, M. (1976). *Gaze and mutual gaze*. London: Cambridge University Press.
- Argyle, M. & Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry*, **28**, 289-304.
- Badin, P., Bailly, G., Revéret, L., Baciú, M., Segebarth, C. & Savariaux, C. (2002). Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images. *Journal of Phonetics*, **30**(3), 533-553.
- Bailenson, J., Blascovich, J., Beall, A. & Loomis, J. (2001). Equilibrium theory revisited: mutual gaze and personal space in immersive virtual environments. *Presence: teleoperators and virtual environments*, **10**(6), 583-597.
- Bailly, G., Béjar, M., Elisei, F. & Odisio, M. (2003). Audiovisual speech synthesis. *International Journal of Speech Technology*, **6**, 331-346.
- Bailly, G., Elisei, F., Badin, P. & Savariaux, C. (2006a). Degrees of freedom of facial movements in face-to-face conversational speech. In *International Workshop on Multimodal Corpora*, vol., pp. 33-36. Genoa - Italy.
- Bailly, G., Elisei, F., Raidt, S., Casari, A. & Picot, A. (2006b). Embodied conversational agents : computing and rendering realistic gaze patterns. In *Pacific Rim Conference on Multimedia Processing*, vol. LNCS 4261, pp. 9-18. Hangzhou.
- Barattelli, S., Sichelschmidt, L. & Rickheit, G. (1998). Eye-movements as an input in human computer interaction: exploiting natural behaviour. In *24th annual conference of the IEEE*, vol. 4, pp. 2000-2005. Aachen, Germany.
- Baron-Cohen, S. (1995). *Mindblindness*. Boston, MA: MIT Press.
- Bechler, D., Schlosser, M.S. & Kroschel, K. (2004, 28 Sept.-2 Oct.). System for robust 3D speaker tracking using microphone array measurements. *Intelligent robots and systems*, pp. 2117- 2122.
- Benoît, C., Grice, M. & Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, **18**, 381-392.
- Béjar, M. (2003). *Mouvements oculaires, accès lexical et deixis*. Grenoble: DEA Sciences Cognitives - INPG.
- Beskow, J. (1995). Rule-based Visual Speech Synthesis. In *Eurospeech*, vol. 1, pp. 299-302. Madrid, Spain.
- Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E. & Öhman, T. (1997). The Teleface project - multimodal speech communication for the hearing impaired. In *Eurospeech*, vol. 4, pp. 2003-2010. Rhodes, Greece.
- Bilvi, M. & Pelachaud, C. (2003). Communicative and statistical eye gaze predictions. In *International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, vol., pp. Melbourne, Australia.
- Bond, E.K. (1972). Perception of form by the human infant. *Psychological bulletin*, **77**, 225-245.
- Breazeal, C. (1998). A motivational system for regulating human-robot interaction. In *fifteenth national conference on artificial intelligence*, vol., pp. 54-61. Madison, WI.
- Bregler, C., Covell, M. & Slaney, M. (1997). VideoRewrite: driving visual speech with audio. In *SIGGRAPH'97*, vol., pp. 353-360. Los Angeles, CA.

- Brooks, R. & Meltzoff, A.N. (2005). The development of gaze following and its relation to language. *Developmental Science*, **8**, 535–543.
- Bruner, J. & Sherwood, V. (1976). Early rule structure: The case of peekaboo. In *Life sentences: Aspects of the social role of language* (R. Harre, editor), London, New York: John Wiley.
- Buisine, S., Abrilian, S. & Martin, J.-C. (2004). Evaluation of multimodal behaviour of embodied agents. In *From brows to trust: evaluating embodied conversational agents* (Z. Ruttkay & C. Pelachaud, editors), pp. 217-238. Kluwer Academic Publishers.
- Burroughs, W., Schultz, W. & Autrey, S. (1973). Quality of argument, leadership votes, and eye contact in three-person leaderless groups. *Journal of social psychology*, **90**(1), 89-93.
- Butterworth, G. (2003). Pointing is the royal road to language for babies. In *Pointing : Where language, culture, and cognition meet* (S. Kita, editor), pp. 9–33. Mahwah, NJ: Lawrence Erlbaum Associates.
- Casari, A. (2006). *Modélisation réaliste et gestion du regard d'un clone parlant dans un environnement de réalité partagée*. Unpublished Master Recherche SIPT, INP, Grenoble.
- Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsón, H. & Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Conference on Human factors in computing systems*, vol., pp. 520 - 527. Pittsburgh, Pennsylvania, United States, ACM New York, NY, USA.
- Cassell, J., Torres, O.E. & Prevost, S. (1999). *Turn taking vs. discourse structure: how best to model multimodal conversation*. The Hague: Kluwer.
- Chen, M. (2002). Leveraging the asymmetric sensitivity of eye contact for videoconference. In *SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, vol., pp. 49-56. Minneapolis, Minnesota.
- Cline, M.G. (1967). The perception of where a person is looking. *American Journal of Psychology*, **80**, 41–50.
- Cohen, M.M. & Massaro, D.W. (1993). Modeling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation* (D. Thalmann & N. Magnenat-Thalmann, editors), pp. 141-155. Tokyo: Springer-Verlag.
- Colburn, R.A., Cohen, M.F. & Drucker, S.M. (2000). *The role of eye gaze in avatar mediated conversational interfaces* (Technical report): Microsoft Research.
- Cook, M. & Smith, J.M.C. (1975). The role of gaze in impression formation. *British journal of social and clinical psychology*, **14**, 19-25.
- Cootes, T.F., Edwards, G.J. & Taylor, C.J. (2001). Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(6), 681-685.
- Cosatto, E. & Graf, H.P. (1997). Sample-based synthesis of photo-realistic talking-heads. In *SIGGRAPH'97*, vol., pp. 353-360. Los Angeles, CA.
- Cosatto, E. & Graf, H.P. (1998). Sample-based of photo-realistic talking heads. In *Computer Animation*, vol., pp. 103-110. Philadelphia, Pennsylvania.
- Coutts, L.M. & Schneider, F.W. (1975). Visual behavior in an unfocused interaction as a function of sex and distance. *Journal of experimental social psychology*, **11**, 64-77.
- Dautenhahn, K., Ogden, B. & Quick, T. (2002). From embodied to socially embedded agents - implications for interaction-aware robots. *Cognitive Systems Research*, **3**(3), 397-428.
- Driver, J., Davis, G., Riccardelli, P., Kidd, P., Maxwell, E. & Baron-Cohen, S. (1999a). Shared attention and the social brain : gaze perception triggers automatic visuospatial orienting in adults. *Visual Cognition*, **6**(5), 509-540.

- Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E. & Baron-Cohen, S. (1999b). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition*, **6**, 509-540.
- Ekman, P. & Friesen, W. (1978). *Facial Action Coding System (FACS): A technique for the measurement of facial action*. Palo Alto, California.: Consulting Psychologists Press.
- Ekman, P. & Friesen, W.V. (1975). *Unmasking the Face*. Palo Alto, California.: Consulting Psychologists Press.
- Engbert, R. & Kliegl, R. (2003). Binocular coordination in microsaccades. In *The mind's eye: cognitive and applied aspects of eye movement research* (J. Hyönä, R. Radach & H. Deubel, editors), pp. 103-118. Amsterdam: Elsevier.
- Evinger, C., Manning, K., Pellegrini, J., Basso, M., Powers, A. & Sibony, P. (1994). Not looking while leaping: the linkage of blinking and saccadic gaze shifts. *Experimental Brain Research*, **100**, 337-344.
- Ezzat, T., Geiger, G. & Poggio, T. (2002). Trainable videorealistic speech animation. *ACM Transactions on Graphics*, **21**(3), 388-398.
- Ezzat, T. & Poggio, T. (1998). MikeTalk: a talking facial display based on morphing visemes. In *Computer Animation*, vol., pp. 96-102. Philadelphia, PA.
- Fagel, S. (2006). Joint Audio-Visual Unit Selection - The JAVUS Speech Synthesizer. In *International Conference on Speech and Computer*, vol., pp. St. Petersburg.
- Fantz, R.L. (1965). Visual perception from birth as shown by pattern selectivity. *Annals of the New York academy of science*, **118**, 793-814.
- Farroni, T., Csibra, G., Simion, F. & Johnson, M.H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(14), 9602-9605.
- Freedman, E.G. & Sparks, D.L. (1997). Eye-head coordination during head-unrestrained gaze shifts in Rhesus monkeys. *Journal of Neurophysiology*, **77**, 2328-2348.
- Friesen, C.K., Ristic, J. & Kingstone, A. (2004). Attentional effects of counterpredictive gaze and arrow cues. *Journal of Experimental Psychology: Human Perception and Performance*, **30**(2), 319-329.
- Gamer, M. & Hecht, H. (2007). Are you looking at me? Measuring the cone of gaze. *Journal of Experimental Psychology: Human Perception and Performance*, **33**(3), 705-715.
- Garau, M., Slater, M., Bee, S. & Sasse, M.A. (2001). The impact of eye gaze on communication using humanoid avatars. In *SIGCHI conference on Human factors in computing systems*, vol., pp. 309-316. Seattle, Washington.
- Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A. & Sasse, M.A. (2003). The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. In *SIGCHI*, vol., pp. 529-536. Lauderdale, Florida, USA.
- Geiger, G., Ezzat, T. & Poggio, T. (2003). *Perceptual evaluation of video-realistic speech* (CBCL Paper #224/AI Memo #2003-003). Cambridge, MA: Massachusetts Institute of Technology.
- Gescheider, G.A. (1985). *Psychophysics; Method, theory, and application*. London: Lawrence Erlbaum Assoc Inc.
- Gibson, J.J. & Pick, A.D. (1963). Perception of another person's looking behavior. *American Journal of Psychology*, **76**(3), 386-394.
- Godijn, R. & Theeuwes, J. (2003). The relationship between exogenous and endogenous saccades and attention. In *The mind's eye: cognitive and applied aspects of eye movement research* (J. Hyönä, R. Radach & H. Deubel, editors), pp. 3-26. Amsterdam: North-Holland.

- Goldin-Meadow, S. & Butcher, C. (2003). Pointing toward two-word speech in young children. In *Pointing: Where language, culture, and cognition meet* (S. Kita, editor), pp. 85-109. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gullberg, M. & Holmqvist, K. (2001). Visual attention towards gestures in face-to-face interaction vs on screen. In *International Gesture Workshop*, vol., pp. 206-214. London, UK.
- Hall, E.T. (1963). A System for the Notation of Proxemic Behaviour. *American Anthropologist*, **85**, 1003–1026.
- Harris, C.S., Thackray, R.I. & Schoenberger, R.W. (1966). Blink rate as a function of induced muscular tension and manifest anxiety. *Perceptual motor skills*, **22**, 155-160.
- Hellwig, B. & Uytvanck, D. (2004). *EUDICO Linguistic Annotator (ELAN) Version 2.0.2 manual*. Nijmegen - NL: Max Planck Institute for Psycholinguistics.
- Hess, E.H. (1965). Attitude and pupil size. *Scientific american*, **212**, 46-54.
- Heylen, D.K.J., Es, I., Nijholt, A. & van Dijk, E.M.A.G. (2005). *Controlling the gaze of conversational agents*: Kluwer Academic Publishers.
- Itti, L., Dhavale, N. & Pighin, F. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. In *SPIE 48th Annual International Symposium on Optical Science and Technology*, vol., pp. 64-78. San Diego, CA.
- Jacob, R.J.K. (1993). Eye Movement-Based Human-Computer Interaction Techniques: Toward Non-Command Interfaces. *Advances in Human-Computer Interaction*, **4**, 151-190.
- Johnson, M.H. & Morton, J. (1991). *Biology and Cognitive Development: The Case of Face Recognition*. Oxford: Blackwell.
- Kampe, K.K.W., Frith, C.D. & Frith, U. (2003). "Hey John": Signals Conveying Communicative Intention toward the Self Activate Brain Regions Associated with "Mentalizing," Regardless of Modality. *The Journal of Neuroscience*, **23**(12), 5258-5263.
- Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, **26**, 22-63.
- Klaus, M.H., Jerauld, R., Kreger, N., McAlpine, W., Steffa, M. & Kennell, J.H. (1972). Maternal attachment: Importance of the first post-partum days. *New England Journal of Medicine*, **286**, 460-463.
- Kleck, R. & W. Nuessle, B.J.S.C.P., pp. . (1968). Congruence between the indicative and communicative functions of eye contact in interpersonal relations. *The British journal of social and clinical psychology*, **7**(4), 241–246.
- Kleinke, C.L. (1986). Gaze and eye contact: a research review. *Psychological Bulletin*, **100**(1), 78-100.
- Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thorisson, K. & Vilhjalmsson, H. (2006). Towards a common framework for multimodal generation in ECAs: The behavior markup language. In *Intelligent Virtual Agents* (J.G.e. al., editor, vol., pp. 205-217. Marina del Rey, Springer-Verlag, Berlin.
- Langton, S., Watt, J. & Bruce, V. (2000). Do the eyes have it ? Cues to the direction of social attention. *Trends in Cognitive Sciences*, **4**(2), 50-59.
- Lansing, C.R. & McConkie, G.W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research*, **42**(3), 526-539.
- Lee, S.P., Badler, J.B. & Badler, N. (2002). Eyes alive. *ACM Transaction on Graphics*, **21**(3), 637-644.
- Liversedge, S.P. & Findlay, J.M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, **4**(1), 6-14.

- Magnusson, M.S. (2000). Discovering hidden time patterns in behavior: T-appterns and their detection. *Behavior Research Methods, Instruments, & Computers*, **32**, 93-110.
- Masataka, N. (2003). From index-finger extension to index-finger pointing: ontogenesis of pointing in preverbal infants. In *Pointing: Where language, culture, and cognition meet* (S. Kita, editor), pp. 69-84. Mahwah, NJ: Lawrence Erlbaum Associates.
- Massaro, D.W., Cohen, M.M. & Beskow, J. (2000). Developing and evaluating conversational agents. In *Embodied conversational agents* (J. Cassell, J. Sullivan, S. Prevost & E. Churchill, editors), pp. 287-318. Cambridge, MA: MIT Press.
- Matusaka, Y., Fujie, S. & Kobayashi, T. (2001). Modeling of conversational strategy for the robot participating in the group conversation. In *7th european conference on speech communication and technology*, vol., pp. Aalborg, Denmark.
- McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.
- Minato, T., Shimada, M., Itakura, S., Lee, K. & Ishiguro, H. (2005). Does gaze reveal the human likeness of an android? *4th IEEE International Conference on Development and Learning*, 106-111.
- Mukawa, N., Oka, T., Arai, K. & Yuasa, M. (2005). What is connected by mutual gaze? - user's behavior in video-mediated communication. In *Conference on human factors in computing systems*, vol., pp. 1677 - 1680. Portland, OR, USA, ACM New York, NY, USA.
- Munhall, K., Jones, J.A., Callan, D.E., Kuratate, T. & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. *Psychological Science*, **15**(2), 133-137.
- Nickel, K. & Stiefelhagen, R. (2003). Recognition of 3D-pointing gestures for human-robot-interaction. In *Humanoids*, vol., pp. Karlsruhe, Germany.
- Noll, A.M. (1976). The effects of visible eye and head turn on the perception of being looked at. *The american journal of psychology*, **89**(4), 631-644.
- Novick, D.G., Hansen, B. & Ward, K. (1996). Coordinating turn-taking with gaze. In *ICSLP*, vol. 3, pp. 1888-1891. Philadelphia, PA.
- Paré, M., Richler, R., ten Hove, M. & Munhall, K.G. (2003). Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect. *Perception and Psychophysics*, **65**, 553-567.
- Pelachaud, C. (2002). Visual text-to-speech. In *MPEG4 Facial Animation - The standard, implementations and applications* (I.S. Pandzic & R. Forchheimer, editors), pp. 125-140. John Wiley & Sons.
- Peters, C. (2006). A perceptually-based theory of mind model for agent interaction initiation. *International Journal of Humanoid Robotics*, **3**(3), 321 - 340.
- Peters, C. & O'Sullivan, C. (2003). Bottom-up visual attention for virtual human animation. In *Computer Animation and Social Agents*, vol., pp. 111-117. Rutgers University, New York.
- Picot, A., Bailly, G., Elisei, F. & Raidt, S. (2007). Scrutinizing natural scenes: controlling the gaze of an embodied conversational agent. In *International Conference on Intelligent Virtual Agents (IVA)*, vol., pp. 272-282. Paris.
- Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V. & de Carolis, B. (2005). GRETA. A believable embodied conversational agent. In *Multimodal intelligent information presentation* (O. Stock & M. Zancarano, editors), pp. 3-26. Dordrecht: Kluwer.
- Ponder, E. & Kennedy, W.P. (1927). On the act of blinking. *Quarterly Journal of Experimental Physiology* **18**(2), 89-110.
- Posner, M.I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, **32**, 3-25.
- Pritchard, R.M., Heron, W. & Hebb, D.O. (1960). Visual perception approached by the method of stabilised images. *Canadian Journal of Psychology*, **14**, 67-77.

- Rabiner, L.R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *IEEE Trans. on Audio, Speech, and Language Processing*, **77**, 257-286.
- Ramat, S., Schmid, R. & Zambardi, D. (2003). Eye-head coordination in darkness: Formulation and testing of a mathematical model. *Journal of Vestibular Research*, **13**, 79-91.
- Ricciardelli, P., Baylis, G. & Driver, J. (2000). The positive and negative of human expertise in gaze perception. *Cognition*, **77**(1), B1-B14.
- Robinson, D.A. (1968). The oculomotor control system: a review. In *inst. electrical & electronic engineers*, vol. 56, pp. 1032-1049.
- Ross, M., Layton, B., Erickson, B. & Schopler, J. (1973). Affect, facial regard, and reactions to crowding. *Journal of Personality and Social Psychology*, **28**(1), 69-76.
- Salvucci, D.D. & Goldberg, J.H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Eye Tracking Research and Applications Symposium*, vol., pp. 71-78. Palm Beach Gardens, FL.
- Santella, A. & DeCarlo, D. (2004). Robust clustering of eye movement recordings for quantification of visual interest. In *Symposium on eye tracking research & applications*, vol., pp. 27 - 34. San Antonio, Texas, New York, NY, USA.
- Scassellati, B. (2001). *Foundations for a theory of mind for a humanoid robot*. MIT, Boston - MA.
- Simons, D.J. & Chabris, C.F. (1999). Gorillas in our midst: sustained inattention blindness for dynamic events. *Perception*, **28**(9), 1059-1074.
- Sparks, D.L. (2002). The brainstem control of saccadic eye movements. *Nature Reviews Neuroscience*, **3**, 952-964.
- Spitz, R.A. & Wolf, K.M. (1946). The smiling response : a contribution to the ontogenesis of social relations. *Genetic Psychology Monographs*, **34**, 57-125.
- Stiefelhagen, R., Yang, J. & Waibel, A. (1997). Tracking Eyes and Monitoring Eye Gaze. In *Workshop on Perceptual User Interfaces*, vol., pp. 98-100. Canada.
- Surakka, V., Illi, M. & Isokoski, P. (2003). Voluntary eye movements in human-computer interaction. In *The mind's eye: cognitive and applied aspects of eye movement research* (J. Hyönä, R. Radach & H. Deubel, editors), pp. 473-491. Amsterdam, The Netherlands: North-Holland.
- Svanfeldt, G., Wik, P. & Nordenberg, M. (2005). Artificial gaze - perception experiment of eye gaze in synthetic faces. In *Second Nordic Conference on Multimodal Communication*, vol., pp.
- Symons, L.A., Hains, S.M.J. & Muir, D.W. (1998). Look at me: five-month-old infants' sensitivity to very small deviations in eye-gaze during social interactions. *Infant Behavior and Development*, **21**(3), 531-536.
- Symons, L.A., Lee, K., Cedrone, C.C. & Nishimura, M. (2004). What are you looking at? Acuity for triadic eye gaze. *The Journal of general psychology*, **131**(4), 451-469.
- Theobald, B.J., Bangham, J.A., Matthews, I. & Cawley, G.C. (2001). Visual speech synthesis using statistical models of shape and appearance. In *Auditory-Visual Speech Processing Workshop*, vol., pp. 78-83. Scheelsminde - Denmark.
- Thórisson, K. (2002). Natural turn-taking needs no manual: computational theory and model from perception to action. In *Multimodality in language and speech systems* (B. Granström, D. House & I. Karlsson, editors), pp. 173-207. Dordrecht, The Netherlands: Kluwer Academic.
- Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, **59**, 433-460.
- Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S. & Munhall, K.G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, **60**, 926-940.

- Vatikiotis-Bateson, E., Munhall, K.G., Hirayama, M., Kasahara, Y. & Yehia, H. (1996). Physiology-based synthesis of audiovisual speech. In *Speech Production Seminar: Models and Data*, vol., pp. 241-244. Autrans, France.
- Vertegaal, R., Slagter, R., van der Veer, G. & Nijholt, A. (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Conference on Human Factors in Computing Systems*, vol., pp. 301 - 308. Seattle, USA, ACM Press New York, NY, USA.
- Vinayagamorthy, V., Garau, M., Steed, A. & Slater, M. (2004). An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience. *The Computer Graphics Forum*, **23**(1), 1-11.
- Wallis, P. (2005). Believable conversational agents: Introducing the intention map. In *International conference on autonomous agents and multiagent systems*, vol., pp. 17-22. Utrecht.
- Weiss, C. (2004). Framework for data-driven video-realistic audio-visual speech synthesis. In *Int. Conf. on Language Resources and Evaluation*, vol., pp. Lisbon.
- Yarbus, A.L. (1967). Eye movements during perception of complex objects. In *Eye Movements and Vision* (L.A. Riggs, editor), pp. 171-196. New York: Plenum Press.

7 RESUME EN FRANÇAIS DE LA THESE

Gaze and face-to-face communication between a human speaker and an animated conversational agent – Mutual attention and multimodal deixis

Regard et communication face-à-face entre un locuteur humain et un agent conversationnel animé. Attention mutuelle et monstration multimodale

présentée et soutenue publiquement par

Stephan Raidt

le 2 Avril 2008

DIRECTEUR DE THESE

Gérard Bailly

CO-DIRECTEUR DE THESE

Laurent Bonnaud

7.1 INTRODUCTION

Le sujet de cette thèse fait partie de l'objectif visé par notre laboratoire d'analyser et modéliser le protocole d'échange d'information linguistique et paralinguistique dans la conversation humaine. Dans ce contexte nous nous intéressons aux signaux multimodaux qui encodent ces informations avec le but de déterminer quand, comment et pourquoi ces informations doivent être traitées.

La structure de la parole acoustique et notamment de sa prosodie est déjà bien maîtrisée. Ces informations aident les interlocuteurs à réguler le flux d'information, à subdiviser le flux audio en segments pertinents et à communiquer des attitudes et émotions. D'autres signaux multimodaux plus orientés vers une émergence visuelle contribuent également à la structuration de ces informations.

Les gestes du corps, de la main, de la tête, du visage et des yeux donnent en continu à nos interlocuteurs des informations sur l'attitude et les émotions en relation avec les énoncés échangés. Ces gestes sont tellement fortement intégrés dans notre manière de communiquer, que nous les produisons même si le destinataire de nos informations n'est physiquement pas présent (e.g. conversation téléphonique). Si, dans ces cas, les informations gestuelles n'apportent pas de gain d'information au récipiendaire, ils aident toujours l'émetteur à réduire sa charge cognitive, notamment lors de la planification de son propre discours. Ces gestes font partie de notre comportement sensi-moteur et ces activités motrices aident dans le déclenchement approprié de nos processus cognitifs. Nous sommes incapables de simplement arrêter de nous comporter comme des êtres sociaux, même lorsque l'interlocuteur n'est pas physiquement présent.

Le travail présenté dans ce rapport est dédié à l'étude du regard et en particulier à l'étude de son rôle dans la direction de l'attention de nos interlocuteurs vers des objets dans l'environnement et son apport à maintenir leur attention sur notre discours. Dans l'interaction face-à-face, le regard a de multiples fonctions notamment en relation avec la structuration du dialogue et de l'argumentation. Le regard est également une composante importante dans la

régulation des tours de parole, la prise de la parole, la génération d'attention mutuelle et de l'interaction fondée. Souvent, dans ce contexte, les capacités déictiques du regard et la régulation de l'attention mutuelle par le regard sont concomitantes. La plupart des activités sociales incluent des tâches collaboratrices pour lesquelles l'environnement joue une grande importance dans la conversation face-à-face.

Pour les caractères virtuels, ces fonctions sont cruciales pour pouvoir générer l'impression de présence et de conscience de l'environnement. Pour les robots, leur simple présence physique suffit partiellement à ce propos, puisqu'ils peuvent démontrer de manière directe une conscience de l'environnement par la manipulation des objets ou en évitant des obstacles. Les caractères virtuels affichés sur un écran ont besoin de plus d'effort pour prouver les mêmes capacités. Une animation cohérente du regard ainsi que des réactions appropriées au comportement d'un interlocuteur humain sont des moyens nécessaires et performants pour générer l'impression de présence et de conscience. Cela nécessite évidemment une riche analyse de scène ainsi que des modèles d'interaction appropriés.

L'analyse et la modélisation du comportement de regard impliquent des boucles de perception-action de différentes portées. La poursuite continue d'un objet en mouvement nécessite un couplage de bas-niveau entre la perception rétinienne et le contrôle du regard. L'attention vers des éléments référencés dans le discours et présents dans la scène, nécessite souvent de l'analyse, de la compréhension et de la synthèse des actes de dialogue d'une portée temporelle correspondant à la longueur de plusieurs énoncés.

Le but final de notre travail est la modélisation du regard pour un agent conversationnel animé. Notre méthodologie dans cette approche est plutôt de nature expérimentale. Nous avons étudié et caractérisé l'interaction dyadique entre deux personnes en temps réel ainsi que l'interaction entre un humain et un agent animé. Dans ce contexte, nos objectifs sont à la fois la précision et la pertinence. La précision est assurée par la mesure précise de la direction du regard durant l'interaction par des techniques d'oculométrie actuelle. Dans cette thèse, nous présentons l'analyse détaillée de caractéristiques multimodales incluant le regard, la parole et les mouvements faciaux en tant que signaux de bas niveaux ainsi que des informations de plus haut niveau concernant les activités cognitives des interlocuteurs qui représentent la motivation des signaux mesurés. La pertinence est assurée par la conception de scénarios qui ont comme objectif de contrôler l'interaction et de faciliter l'interprétation des comportements observés. Ces travaux sont donc très spécialisés et ne cherchent pas à proposer une description générale de toutes les stratégies possibles pour l'implémentation des capacités de regard d'un agent artificiel. Ils présentent néanmoins une analyse et modélisation très détaillées du regard observé pendant des situations d'interaction bien contrôlées orientées vers des tâches. Le rapport est structuré en quatre parties principales :

- Chapitre 1 - 'State of the art' – présente un résumé des recherches sur le regard et la perception visuelle qui sont importantes dans le contexte de nos travaux d'expérimentation ainsi que pour la compréhension générale. Nous présentons également un résumé concernant les agents conversationnels animés qui sont la base de notre plateforme expérimentale.

Le travail expérimental est séparé en deux parties.

- Chapitre 2 - 'Gaze direction and attention' – décrit deux expérimentations dédiées à l'analyse de l'impact de gestes déictiques donnés par une tête parlante sur la performance des usagers dans une simple tâche de recherche et sélection sur un écran d'ordinateur. Nous avons testé à quel point notre tête parlante est capable de générer des gestes déictiques convaincants et comment un usager peut en profiter. Les résultats de ces expériences montrent que, dans une tâche de recherche et sélection, les indices corrects donnés par la tête parlante réduisent le temps de réaction des sujets. Ces gestes peuvent

aussi réduire le nombre d'objets inspectés pendant la recherche, ce que nous interprétons comme une réduction de la charge cognitive. Il apparaît par contre que les sujets arrivent à ignorer les faux indices, ce qui suggère que l'impact de la présence de la tête parlante et de ses gestes déictiques ne domine pas les stratégies conscientes de recherche des sujets. L'amélioration de la performance par l'accompagnement des gestes de la tête et de la main par des énoncés impératifs est un résultat très intéressant. Cette accentuation de l'instant auquel l'information gestuelle devrait être interprétée paraît être un facteur très important dans la direction d'attention. Le geste multimodal spécifiant l'endroit par le geste visuel et l'instant de son interprétation par la voie auditive montre un impact très fort sur la performance des sujets.

- La deuxième partie expérimentale dans le chapitre 3 - 'Mediated face-to-face interaction' - est un premier essai pour étudier en détail le regard des interlocuteurs dans une interaction face-à-face. Les travaux les plus importants dans la recherche du regard dans le contexte de l'interaction humaine datent des années 1970. A l'époque, les techniques oculométriques étaient beaucoup moins développées et en général se basaient sur les observations des expérimentateurs. La résolution temporelle des mesures ainsi que la précision angulaire de la direction du regard étaient relativement faibles. Depuis, les connaissances sur le regard dans l'interaction face-à-face ont peu évoluées. Nos travaux consacrés à l'interaction face-à-face médiatisée reprennent les recherches dans ce domaine, avec la perspective d'améliorer le contrôle du regard des caractères virtuels. Ces travaux incluent le développement et l'implémentation d'une plate-forme d'expérimentation et d'un scénario approprié ainsi que l'exécution d'une série d'expériences basées sur des interactions dyades orientées vers des tâches. A partir de l'analyse des données expérimentales, nous avons développé un modèle pour l'animation du regard d'un agent conversationnel animé. Les résultats confirment que les yeux et la bouche sont des cibles de fixation saillantes, comme c'était déjà postulé dans les travaux d'autres chercheurs. Nous montrons que le comportement du regard est influencé par les différents états de la conversation que nous distinguons. Comme attendu, il y a des variations entre les différents individus. Notre sujet cible par contre montre un comportement relativement consistant et une répartition bien équilibrée sur les yeux et la bouche comme cibles de fixation. Dans certains cas, on observe des relations très prononcées entre les segments de l'interaction et le comportement du regard. Notre sujet cible par exemple montre une forte tendance à regarder la bouche pendant l'écoute, ce qui est encore intensifié quand il attend des informations nouvelles. Les yeux restent des cibles saillantes pendant toute l'expérience et sont regardés régulièrement. Nous n'avons pas pu trouver des relations entre le comportement du regard et l'occurrence de regard mutuel dans nos données. Ni un évitement, ni une recherche de regard mutuel n'ont pu être montrés avec les données mesurées. Quand le sujet cible est remplacé par une vidéo pré enregistrée avec le but de mettre les sujets en interaction, cela entraîne un changement du comportement du regard dans certains cas. Les observations suggèrent que ces variations sont liées au degré avec lequel un sujet est convaincu d'interagir avec une vraie personne. L'analyse du clignement de l'œil montre également des relations entre les segments de l'interaction et la fréquence des clignements. Les observations donnent lieu à l'hypothèse qu'une augmentation de l'attention entraîne une réduction de la fréquence des clignements des yeux. En plus, l'analyse montre une tendance remarquable de notre sujet cible à exécuter des clignements pendant qu'il se prépare à parler, surtout en cooccurrence avec un mouvement majeur de la tête. Cela pourrait avoir comme raison des réflexes protecteurs ou bien fonctionner comme un signal relié à la prise de parole, similaire à l'aversion du regard observé par d'autres chercheurs au début d'un tour de parole.

Le chapitre 4 - 'Conclusions' – résume les résultats que nous avons obtenus et discute les perspectives pour des travaux à venir et la thématique de cette thèse.

7.2 ÉTAT DE L'ART – REGARD ET PERCEPTION VISUELLE

Le regard est l'élément le plus important analysé dans cette thèse. Ce chapitre résume les bases de la perception visuelle humaine et ses rapports avec la communication et l'interaction humaines. Nous expliquons la physiologie de l'œil ainsi que les fonctions de son système moteur.

La connaissance des caractéristiques de l'œil est essentielle pour la compréhension de son apparence visuelle et le rôle dans la perception visuelle de celle-ci. Pour permettre de correctement apprécier les citations de la littérature, nous décrivons les différentes techniques d'oculométrie utilisées dans les recherches citées. Ensuite, nous expliquons l'importance cruciale de l'œil et de l'orientation du regard comme stimuli visuels et comment elle émerge de l'importance du sens visuel pour la perception humaine. Les fonctions principales que le regard peut jouer dans l'interaction humaine seront également discutées.

7.2.1 La physiologie de l'œil

Cette section décrit en détail la configuration de l'œil et son système moteur. Selon les caractéristiques des mouvements de l'œil, les traitements cognitifs et le contexte dans lequel les mouvements sont observés, différents types de mouvements sont distingués. Dans cette discussion nous incluons les clignements des yeux comme une activité liée aux activités de l'œil.

7.2.2 Mesure de la direction du regard

Pour mesurer la direction du regard il existe des méthodes variées. Avec l'évolution des connaissances techniques les méthodes utilisées dans les expériences scientifiques ont également évoluées avec le temps. Les conditions expérimentales définies par la tâche et l'installation expérimentale et spécialement la méthode appliquée pour la mesure du regard sont des critères importants pour bien savoir interpréter les données acquises ainsi que les résultats d'évaluation.

A ce propos nous décrivons les différentes méthodes utilisées par les chercheurs dans les travaux cités.

7.2.3 Perception visuelle des yeux et du regard

Les yeux et le regard d'une autre personne sont des stimuli très particuliers dans toutes les cultures connues. Même chez les animaux la perception des yeux déclenche des réactions et des comportements particuliers.

Ce chapitre discute en détail comment la sensibilité de la perception visuelle aux yeux se développe chez les humains, la précisons avec laquelle la direction du regard peut être estimée et comment le regard d'une autre personne est interprété comme geste déictique qui peut déclencher des changements d'attention.

7.2.4 Le rôle du regard dans l'interaction sociale

Le regard et les yeux sont connus pour avoir une signification très importante dans la perception humaine. C'est une suite naturelle qu'ils jouent un rôle important dans la communication et dans l'interaction humaine. Butterworth (2003) discute l'importance de

pointage en combinaison avec la parole dans le développement de la capacité à référencer des objets par des moyens linguistiques. Il explique également comment le pointage, la parole et le regard sont liés pour établir de l'attention visuelle et leur importance dans une communication bien fondée. Même à travers de cultures et de barrières langagières, ces aspects du regard sont d'une intelligibilité générale. Il est évident que le regard prend un rôle exceptionnel entre les différentes modalités de communication. Il est capable de communiquer des contenus est des informations ainsi que d'influencer ou de modifier le contenu transmit par la parole ou d'autre modalités de communication.

Ce chapitre discute le regard dans le contexte de l'interaction sociale. Plusieurs chercheurs faisaient des efforts à étudier, à décrire et à quantifier les valeurs communicatives du regard et la grande variété de fonctions qu'il peut jouer.

7.2.5 Le regard comme modalité de commande dans l'interaction homme-machine

Dans les implémentations actuelles utilisant le regard humain comme modalité de commande dans l'interaction home-machine, l'utilisation du regard est mieux décrite en la comparant avec la souris d'ordinateur et son utilisation pour sélectionner des objets.

Pour une telle utilisation du regard comme moyen de sélectionner des objets sur un écran d'ordinateur, différentes stratégies sont connues. Un simple critère pour la sélection est le temps de fixation, mesuré à l'aide d'un oculomètre. Dans ce cas, un seuil minimal de durée est défini pour déclencher la sélection d'un objet aussitôt que la durée du regard vers cette objet dépasse le seuil définit.

Indépendamment du critère, une telle utilisation du regard comme modalité d'entrée ne pourrait jamais être une manière d'utilisation naturelle, mais demanderait plutôt de l'apprentissage pour bien pouvoir l'utiliser. L'exploit de comportement naturel du regard dans la sélection des objets nécessiterait la multi-modalité des signaux d'entre mesurés. Cela pourrait par exemple être réalisé par la prise en compte de la parole, qui pourrait servir à préciser l'instant dans lequel l'information sur le lieu extrait de la direction du regard doit être traité pour sélectionner l'objet regardé.

7.2.6 Agents conversationnels animés

Les travaux étant sujet de la thèse présentée exploitent un agent conversationnel animé comme plateforme de recherche et d'expérimentation. Nous informons donc dans ce chapitre de la signification et de la provenance du terme ainsi que des différentes techniques et objectifs de réalisation.

7.3 ATTENTION ET DIRECTION DU REGARD

Le sujet de ces expériences est l'étude de l'impact des gestes déictiques de notre tête parlante sur la performance des sujets dans l'exécution de simples tâches de recherche et de sélection.

Le regard est reconnu comme étant une composante importante de la deixis humaine. Un agent conversationnel animé devrait être équipé d'une telle capacité pour compléter ses fonctionnalités ainsi que pour disposer de moyens nécessaires à l'établissement d'une interaction bien fondée avec un partenaire humain. Les expériences suivantes sont consacrées à la vérification des capacités de notre tête parlante à générer des gestes déictiques convaincants et aux avantages potentiels qu'il est possible d'en tirer.

7.3.1 Montage expérimental et scénario d'interaction

Pour le développement d'un montage expérimental approprié, nous avons suivi les expériences de Langton *et al.* (2000) and Driver *et al.* (1999a) basées sur le paradigme de

Posner. Nous avons copié l'idée principale consistant à utiliser l'image d'une tête humaine pour influencer l'attention d'un usager en l'attirant vers différents points de l'écran.

Contrairement aux expériences suivant le paradigme de Posner, la tête sera présente sur l'écran en permanence pendant l'expérience et d'importance cruciale pour la scène présentée ainsi que la tâche donnée aux sujets. Aussi la réaction demandée au sujet sera plus complexe que la simple signalisation de la perception d'un objet sur l'écran par pression d'un bouton.

Comme scénario, nous avons choisi un jeu de cartes virtuel pour réaliser une tâche de recherche et de sélection. Cela devrait nous permettre de tester à quel point notre tête parlante dispose de moyens de diriger l'attention d'un sujet dans une scène virtuelle complexe par son apparence.

Dans la scène présentée à l'écran, des cartes colorées sont affichées sur les deux bords latéraux. Il y a quatre cartes sur chaque côté, sur lesquelles sont distribués de manière aléatoire à chaque cycle des chiffres compris entre un et huit. Au début de chaque cycle, une carte blanche apparaît en bas au milieu de l'écran. Cette carte, selon le chiffre qu'elle porte, doit être posée sur la carte colorée correspondante à l'aide de la souris. Lorsque le sujet sélectionne la carte blanche en cliquant à l'aide de la souris, les numéros s'affichent sur les cartes latérales.

Pour estimer la performance des sujets, nous surveillons la direction de leur regard ainsi que leur temps de réaction, mesurée entre le click sur la carte blanche qui démarre le cycle, et sa pose sur une carte colorée également par un click de souris. En comparant la direction du regard avec les coordonnées des cartes, nous déterminons combien de cartes ont été regardées par le sujet pendant sa recherche de la carte cible. Un questionnaire donné aux sujets après l'expérience les interroge sur l'estimation subjective de la qualité de la tête parlante et de l'assistance donnée par elle.

Au milieu de l'écran, notre tête parlante peut être affichée optionnellement. Lorsqu'elle l'est, elle peut soit donner des indices corrects, soit des indices faux. Les gestes déictiques sont des gestes composés des mouvements de la tête et des yeux visant un objet sur l'écran. L'expérience est divisée en plusieurs blocs de cycles réalisés dans les mêmes conditions. Les sujets sont informés des conditions et du comportement de la tête parlante si elle est présente dans la scène. En comparant les résultats obtenus dans les différentes conditions expérimentales, nous estimons l'impact de la tête parlant dans cette tâche de recherche et de sélection.

7.3.2 Expérience I: Impact des gestes faciaux

L'expérience I est présentée comme une suite de quatre conditions différentes de 24 cycles chacune. La suite de conditions est maintenue pour tous les sujets. Dans la première condition, la tête parlante n'est pas affichée. L'objectif dans ce cas est de disposer d'une référence de performance à laquelle on peut comparer les résultats d'autres conditions expérimentales. Dans les conditions suivantes la tête est présente. Dans la deuxième condition, elle donne de faux indices, ensuite les indices donnés sont corrects et à la fin vient une condition avec des indices corrects, mais sans que les numéros soient affichés sur les cartes. Cela rend la tâche plus compliquée car il est impossible de trouver la carte cible sans respecter les indices de la tête parlante. Comme l'information supplémentaire au geste déictique présentée sous forme de chiffres affichés n'est plus donnée, seule l'interprétation de la direction du regard permet de repérer la carte recherchée. Cette condition devrait donc donner des résultats qui permettent d'estimer la précision avec laquelle la direction du regard de notre tête parlante peut être interprétée.

En plus de gestes déictiques, la tête énonce des phrases d'encouragement et de félicitation entre les cycles. Ces phrases sont synthétisées à l'avance afin de ne pas surcharger la machine. Selon les différentes conditions, des différences de performance par rapport à la condition

sans la tête parlante peuvent être observées. Quand les indices donnés sont corrects, certains sujets améliorent leur performance. On observe des temps de réaction raccourcis ainsi qu'une réduction du nombre de cartes inspectées pendant la recherche de la carte cible. On s'attendait à observer l'effet inverse lorsque les indices donnés étaient faux. Les résultats indiquent cependant que cet effet n'est pas prononcé. Au contraire, les résultats portent plutôt à croire que les sujets parviennent à ignorer les faux indices donnés par la tête parlante.

Dans le cas d'indices corrects, mais sans l'option de vérifier le choix par comparaison de chiffres (car ces derniers ne sont pas affichés), on observe plusieurs erreurs de sélection. Dans la plupart des cas, il s'agit de confusions avec une carte voisine, ce qui indique un manque de précision du geste déictique.

L'analyse des questionnaires montre une estimation moyenne des qualités de la tête parlante. Nous avons découvert que souvent, les estimations subjectives ne correspondent pas aux mesures objectives de temps de réaction et de nombre de cartes inspectées. On constate également des désaccords entre les résultats objectifs. En effet, un gain de temps n'est pas forcément lié à une réduction du nombre de cartes inspectées.

7.3.3 Expérience II: Impact des gestes déictiques multimodaux

Sur la base des résultats observés pendant la première expérience, les conditions expérimentales ont été modifiées pour une deuxième série d'expériences.

Les phrases d'encouragement et de félicitation provoquaient des réactions spontanées chez les sujets, qui critiquaient le caractère répétitif et aléatoire des propos qui les gênaient. Ces énoncés ont donc été supprimés dans la deuxième expérience. A leur place, la voix a été utilisée pour accentuer le geste déictique. L'énoncé «la» est produit en relation avec le geste déictique. D'un côté, cela devrait renforcer la capacité du geste à attirer l'attention des sujets. D'un autre côté, on spécifie ainsi le temps, quand exactement l'information indiquant le lieu devrait être interprétée.

Nous avons réalisé deux alignements temporels différents entre le geste visuel et l'énoncé vocal. Dans une des conditions, le rapport entre geste et parole est choisi afin d'être perçu comme réaliste. Le même alignement est utilisé pour la condition dans laquelle les indices sont trompeurs, reproduisant ainsi la condition de la première expérience. Un deuxième alignement entre geste et parole induit un retard de la parole à des fins de comparaison. La condition sans assistance, dans laquelle la tête parlante n'est pas affichée, est également reproduite.

Cette fois-ci, l'ordre des quatre conditions est changé entre les sujets afin d'éliminer les effets de l'ordre sur les performances mesurées.

Si l'on compare avec les résultats de la première expérience, les effets observés sont plus prononcés dans les résultats de l'expérience présente. Le nombre de sujets qui profitent d'une réduction significative du temps de réaction ou du nombre de cartes inspectées est augmenté. Aussi les évaluations subjectives données dans le questionnaire sont plus favorables. Néanmoins, tout comme dans la première expérience, il n'y a pas de relation directe entre les différents paramètres mesurés.

7.3.4 Discussion et perspectives

Nous interprétons les résultats des deux expériences comme une confirmation de la capacité de notre tête parlante à diriger avec succès l'attention d'un interlocuteur humain par des gestes multimodaux. D'un côté, cela a pour avantage de réduire le temps de réaction. D'un autre côté, cela facilite la recherche et réduit le nombre d'objets à inspecter avant de trouver l'objet cible. Nous prenons cela comme manifestation d'une charge cognitive réduite par l'assistance proposée.

Le fait que les indices incorrects n'interfèrent pas avec la performance des sujets indique qu'ils réussissent à les ignorer. L'impact de la tête parlante doit donc être estimé comme moins fort qu'attendu étant donné que les sujets sont capables de l'ignorer.

La variation des résultats entre les individus indique qu'ils ont chacun développé des stratégies différentes pour compléter la tâche. Cela était aussi évident dans les commentaires donnés par les sujets après les expériences, qui révélaient des motivations variées.

Un aspect très intéressant est l'amélioration apportée par l'augmentation du geste avec de la parole. Ce geste multimodal semble mieux accepté par les personnes et plus efficace.

En ce qui concerne la poursuite de cette recherche, il y a différentes options. Les problèmes observés lorsqu'il n'y a pas de chiffre sur les cartes a donné lieu à un stage de Master avec pour objectif de développer un modèle détaillé de l'œil et des paupières. Les paupières suivent les mouvements du globe oculaire. Ce sont surtout les mouvements verticaux qui influencent l'apparence des paupières. Les résultats sont très prometteurs et donnent une apparence plus naturelle. En raison d'un manque de temps, ce modèle n'a pas encore pu être testé dans une expérience. Avec le nouveau modèle, nous attendons surtout un gain en précision de l'estimation de la direction du regard.

Pour les futures expériences basées sur ce jeu de cartes, différentes modifications pourraient être intéressantes. Une diversification des objets ainsi qu'une augmentation de leur nombre devraient rendre les différences entre les conditions plus évidentes.

7.4 INTERACTION FACE-A-FACE MEDIATISEE

Les premiers travaux sur le regard dans l'interaction face-à-face distinguent entre deux états conversationnels, 'parler' et 'écouter', ainsi que deux directions de regard, l'une dirigée vers l'interlocuteur, et l'autre non (Kendon (1967), Argyle & Cook (1976)). Comme un événement spécifique, le regard mutuel est rapporté comme l'incident quand les deux personnes interagissant dirigent leur regard vers les yeux de l'autre.

Dans plusieurs études ces catégories sont subdivisées, comme par exemple le début ou la fin du discours. La délimitation des sous-parties n'est par contre pas définie précisément. L'approfondissement de connaissances et les moyens techniques d'aujourd'hui fournissent aux chercheurs les possibilités d'une mesure de la direction du regard beaucoup plus détaillée au niveau temporel ainsi que concernant les angles de vue en relation avec l'orientation de la tête.

L'utilisation de caméras vidéo en plus de l'équipement oculométrique permet d'associer la direction du regard avec des objets ou des cibles dans la scène. Une reprise des recherches dans ce domaine est donc fortement justifiée, avec une segmentation plus détaillée de l'interaction ainsi qu'une mesure du regard plus fine.

Dans la partie de notre travail dédiée à ce sujet nous étudions des conversations face-à-face médiatisées en utilisant des dialogues question-réponse. Les résultats devraient clarifier les relations entre les états des sujets dans une telle interaction et le comportement du regard observé. Nous nous intéressons à ces relations ainsi qu'à l'influence mutuelle qui peut exister entre les comportements du regard de deux sujets interagissant. Dans le contexte de l'animation d'agents conversationnels animés, ces relations sont de grand intérêt. Elles peuvent être utilisées pour les doter des moyens de signaler la conscience de l'environnement, pour ainsi leur donner un air de présence comme base d'une interaction bien fondée et pour signaler de l'attention mutuelle.

7.4.1 Scénario et installation d'expérimentation

L'intérêt principal de ces expériences est l'analyse du comportement du regard pendant l'interaction face-à-face. Les expériences devraient révéler l'influence des activités et des

états mentaux d'un sujet sur son propre comportement de regard. Ces processus endogènes influencent la direction du regard et peuvent ainsi accentuer ou signaler des segmentations de la parole, des gestes de prise de la parole ou peuvent modifier le contenu linguistique. Les résultats devraient aussi clarifier à quel point le regard de l'un influence le regard de l'autre. Le regard résultant des réactions au comportement de l'interlocuteur provient de boucles rapides d'interaction qui jouent un rôle important dans la signalisation d'attention et dans l'établissement d'une interaction fondée.

L'installation d'expérimentation développée pour nos expériences permet de mettre deux sujets en interaction et de les enregistrer en même temps. L'apparence audio-visuelle des sujets ainsi que leurs directions de regard peuvent être enregistrés en fine résolution et d'une manière qui permet de les mettre en relation temporelle et d'associer le regard avec des cibles dans la scène.

Le scénario utilisé facilite la génération récurrente d'événements que nous considérons comme importants dans l'interaction face-à-face. Cela permet de générer un nombre suffisant de données pour des analyses statistiques. Nous avons visé à acquérir un nombre important de données sur un seul sujet, permettant ainsi de développer un modèle de regard imitant le comportement de cette personne. Nous avons choisit comme sujet cible qui interagit avec plusieurs autres sujet la personne qui devait servir comme modèle de notre tête parlant.

7.4.2 Acquisition et traitement de données

Pendant l'expérience, les différentes données sont enregistrées sous différents formats et avec de l'équipement divers. Pour assurer la validité de données et pour pouvoir extraire les paramètres nécessaires pour leur traitement, nous avons développé plusieurs procédures, comme par exemple pour la synchronisation d'enregistrements. Dans ce chapitre, nous détaillons ces procédures ainsi que l'extraction de paramètres nécessaires pour le traitement automatique de données.

7.4.3 Traitement de données

Ici nous décrivons le traitement automatique de données. Il est majoritairement réalisé par de scripts de Matlab. Un script principal fait appel à différentes procédures de Matlab, Perl ou au système d'exploitation pour l'exécution de programmes externes. Le script principal est conçu de manière à séparer la définition de paramètres du traitement de données-mêmes. La définition de paramètres concerne, par exemple, des points de référence pour le suivi des yeux et de la bouche, ou la définition de zones représentant les différentes régions-cible du regard.

7.4.4 Analyse statistique

Pour l'analyse de nos données, nous avons considéré deux approches principales. D'une coté nous cherchions des relations entre les activités et les états internes d'un sujet durant l'interaction et son comportement du regard. Celles-ci sont des processus de contrôle endogènes que nous considérons comme des manifestations conscientes ou inconscientes des états cognitifs. Nous utilisons le terme 'état cognitif' pour spécifier les intervalles de la segmentation de l'interaction. Nous considérons que des activités motrices perceptibles par l'interlocuteur signalent des changements ou la maintenance des ces états cognitifs. En plus, nous distinguons deux rôles que les sujets prennent dans l'interaction, 'initiateur' et 'répondant'. L'initiateur domine le dialogue et donne de nouvelles informations. Le répondant suit plutôt les propos de l'initiateur, reçoit des informations qu'il peut commenter ou il peut demander des précisions. Le rôle est indépendant du tour de parole. Les variables analysées sont la distribution du temps de fixation sur les différentes cibles sur le visage, la probabilité

avec laquelle ces cibles sont regardées, la durée des fixations et l'occurrence des clignements de l'œil.

D'un autre côté nous cherchons des relations entre le regard d'un sujet et le regard de l'autre. Spécialement en cas de contact des yeux, nous nous attendions à observer soit la volonté de maintenir le contact soit de l'éviter. Cela signifierait un contrôle exogène du comportement du regard. Dans ce contexte nous avons étudié l'occurrence du contact des yeux. Nous avons étudié aussi plus généralement le comportement des sujets interagissant autour des événements de regard dirigé directement vers les yeux, indépendamment du contact des yeux. Si cela a un impacte sur le comportement de l'interlocuteur, une réponse moyenne du dernier devrait être observable dans les données spécifiant son direction du regard.

Les analyses confirment avant tout les yeux et la bouche comme des cibles dominant sur le visage. Des analyses MANOVA on montré une influence significative des états cognitifs ainsi que du rôle sur la répartition du temps de fixations sur les différentes cibles. La bouche par exemple devient une cible fréquente seulement dans le rôle répondant, surtout pendant que le sujet écoute. Cela est lié probablement à la recherche des informations supplémentaires par la lecture labiale. Les yeux restent néanmoins des cibles avec une haute probabilité d'être regardés. Ils sont des cibles très fréquentées dans la plupart des états cognitifs et paraissent d'être d'une importance similaire. Seul pendant le court intervalle de la préparation à parler, notre sujet cible à une forte préférence à regarder l'œil droit de son interlocuteur. La durée des fixations varie de manière significative avec la cible de fixation. Des fixations sur le visage ont par exemple la tendance d'être plutôt courtes quand elles ne sont pas dirigées vers les yeux ou la bouche.

L'occurrence des clignements des yeux est également influencée par les états cognitifs et le rôle. De notre analyse émerge la tendance d'une fréquence de clignements d'œil réduit pendant des intervalles d'attention augmentée. Cela se voit clairement par exemple pendant que le sujet cible écoute en tant que répondant.

L'analyse des paramètres de contrôle exogène ne montre pas d'influence sur le comportement du regard. Ni une recherche de contact des yeux ni l'évitement n'est suggéré par les résultats. Probablement cela est dû aux restrictions du scénario. Dans une interaction plus libre, de telles relations serraient sûrement observables.

7.4.5 Modélisation

L'analyse statistique montre l'influence des états cognitifs et du rôle sur le comportement du regard de notre sujet cible. Ces sont des facteurs qui ont un impact sur le temps de fixation, la durée des fixations, et la probabilité d'occurrence des fixations vers les différentes régions d'intérêt aussi que sur l'occurrence de clignements de l'œil.

En nous basant sur ces données, nous avons développé un premier modèle de gestion du regard pour le contrôle du regard de notre agent conversationnel animé. Ce modèle est inspiré par le modèle proposé par Lee *et al.* (2002). Leur approche modélise directement les directions angulaires, les vitesses et les amplitudes de saccades en dépendance de l'état interne du caractère animé. Les fixations ne sont pas dirigées vers des cibles bien définies dans la scène, comme cette information n'était pas prise en compte ni dans le corpus d'entraînement ni dans les processus de génération. Dans ce chapitre nous expliquons comment nous utilisons des HMM pour la génération de fixations, basé sur les résultats de nos expériences ainsi que la génération de clignements de l'œil.

7.5 CONCLUSIONS

Nous avons réalisé deux parties d'expériences indépendantes qui sont toutes les deux dédiées à l'étude des fonctions du regard dans des interactions dialogiques. La première partie

d'expériences utilise une tête parlante déjà existante, capable de produire des mouvements articulatoires et d'orientation de la tête et des yeux. Le but était de vérifier les capacités déictiques de gestes du regard en tant que mouvement joint de tête et des yeux. Comme modification des conditions, nous avons rajouté des commandes de parole concomitantes avec les gestes déictiques. Quand le geste communique en premier les informations décrivant l'endroit, la parole peut spécifier l'instant où cette information devrait être interprétée.

La deuxième partie des expériences est conçue pour la mesure du comportement humain et pour ainsi acquérir des nouvelles connaissances sur l'utilisation du regard pendant une interaction dialogique. Ces connaissances devraient servir à améliorer les capacités communicatives de notre tête parlante par un modèle pertinent pour la gestion du regard. Dans cette perspective, les travaux présentés ici, sont seulement un premier pas vers ces fins. Les interactions que nous avons analysées étaient très restreintes pour réduire la complexité des données ainsi que pour obtenir un nombre suffisant de données qui peuvent être attribuées à des catégories différentes. Les analyses donnent des résultats significatifs et demandent un approfondissement de cette recherche étudiant aussi d'autres aspects de l'interaction communicative afin de pouvoir développer des modèles plus génériques pour la gestion du regard.

7.5.1 Direction du regard et attention

Dans les expériences dédiées à l'étude des capacités déictiques de notre tête parlante, nous avons réalisé différentes conditions d'assistance données par la tête parlante dans un jeu de cartes virtuel. La carte de jeu doit être placée sur la carte cible correspondante à l'aide de la souris d'ordinateur pour sélectionner entre huit cartes cibles possibles.

Nous concluons des résultats que notre tête parlante est capable d'assister les sujets dans cette tâche de recherche et sélection sur écran. Les gestes déictiques fournissent des informations appropriées pour réduire le temps de traitement ainsi que l'effort de la recherche. Même si une réduction du nombre d'objets inspectés pendant cette recherche n'entraîne pas forcément une réduction de temps de réaction, nous considérons cela néanmoins comme un avantage, grâce à une charge cognitive réduite. Le fait que la performance des sujets soit améliorée quand les gestes déictiques sont augmentés par des commandes vocales est une observation très intéressante. Cela est probablement lié à ce rajout de la spécification temporelle de l'instant exact où l'information indiquant l'endroit par les gestes déictiques devrait être interprétée. La parole peut aussi contribuer à attirer encore plus l'attention des sujets vers la tête parlante et l'information donnée par ses gestes en recrutant encore un autre sens de perception. Selon les résultats du questionnaire, cette augmentation de modalités par la parole améliore le jugement de l'apparence de la tête parlante comme plus naturelle.

En somme, les mesures objectives de la performance sont cohérentes avec les évaluations subjectives données dans les questionnaires par rapport à l'amélioration de la tête parlante par l'accompagnement du geste par la parole. Ils ne sont par contre pas nécessairement cohérents considérant les individus séparément. Un sujet peut donner de bonnes évaluations dans le questionnaire mais avoir une mauvaise performance selon les mesures objectives. Même les deux mesures objectives peuvent être contradictoires. Un chemin de recherche raccourci n'entraîne pas nécessairement un temps de réaction réduit par exemple. Cela confirme les observations subjectives de l'expérimentateur que les différents individus ont appliqué des stratégies très différentes pour accomplir la tâche.

Résumant les expériences dédiées à l'étude de la capacité déictique des gestes de notre tête parlante, il y a plusieurs perspectives pour l'amélioration de l'animation ainsi que pour des modifications du scénario expérimental.

Dans la condition dans laquelle les sujets pouvaient seulement retrouver la carte cible en suivant les indices de la tête parlante sans l'option de vérifier par comparaison des chiffres

marqués dessus, plusieurs confusions entre cartes voisines apparaissaient. Cela est probablement dû à l'imprécision de la modélisation des yeux. Une modélisation plus précise des yeux et des paupières pour pouvoir générer des stimuli plus corrects a donné des très bons résultats. Les paupières sont déformées selon l'orientation du globe oculaire en direction horizontale et verticale. Par ce modèle amélioré, une interprétation plus exacte de la direction du regard devrait être possible. La vérification expérimentale de l'impact n'a pas encore eu lieu pour des raisons de manque de temps.

Selon la loi de Fitts, la distance et la taille des objets influencent la performance dans le pointage et dans la sélection d'objets (Surakka *et al.* (2003)). Pour améliorer le scénario expérimental cela devrait être pris en compte, surtout pour le contrôle de l'impact du modèle amélioré de l'œil et des paupières. La diminution de la taille des objets et l'augmentation de leur nombre devraient produire des résultats plus clairs. Désormais, l'impact du geste multimodal pourrait profiter de mesures précises du contrôle moteur du geste et de sa coordination avec la parole.

7.5.2 Interaction face-a-Face médiatisée

Pour l'analyse du comportement du regard dans des situations proches d'un face-à-face, nous avons développé un dispositif expérimental pour une interaction face-à-face médiatisée. Nous avons conduit des expériences avec un scénario limitant les interactions à une conversation de base qui consiste en l'énonciation et la répétition de phrases imposées. Les résultats de ces expériences confirment que ce scénario et le dispositif expérimental sont des moyens appropriés pour analyser le comportement du regard des sujets avec la possibilité de le mettre en relation avec l'interaction en cours. Les états cognitifs et le rôle conversationnel que nous distinguons ont prouvé qu'ils avaient une influence forte sur le comportement du regard et sur les clignements des yeux. L'analyse des données récoltées a permis de développer un modèle de base pour l'animation du regard d'un agent conversationnel animé dans des situations d'interaction face-à-face.

La répartition des fixations confirme clairement les résultats rapportés dans la littérature, qui mentionnent les yeux et la bouche comme des cibles importantes (Vatikiotis-Bateson *et al.* (1998), Lansing & McConkie (1999)). On observe des variations dans le comportement individuel des différents sujets. Notre sujet cible par contre a montré une répartition très cohérente et équilibrée du regard sur les yeux et la bouche pendant les neuf interactions analysées auxquelles elle a participé. Dans certains des cas, il y a une influence très évidente des états cognitifs et du rôle sur ces parcours. Notre sujet cible montre une tendance à regarder la bouche pendant qu'elle écoute qui devient très prononcée dans le rôle répondant où la personne qui écoute reçoit des informations nouvelles. Cela confirme l'argumentation de Lansing & McConkie (1999) que l'attention sur le contenu des mots produit plus de fixations dirigées vers la bouche. Mais même dans les cas où la bouche est une source d'informations importante, les yeux restent toujours des cibles très importants et sont regardés avec une probabilité élevée. Cela accentue l'importance de l'aspect social du comportement du regard pour lequel le regard dirigé vers les yeux a une signification spéciale.

La recherche de relations entre les comportements du regard des deux sujets qui interagissent ne produisait pas des données exploitables. Il n'y avait pas de relation entre le comportement du regard et des événements de regard mutuel. Nous avons attendu soit une recherche soit un évitement du contact visuel. Le fait que nous n'avons pas trouvé de telles relations pourrait être dû au scénario et à l'utilisation de phrases imposées ce qui restreint le contenu de la communication ainsi que sa structure. Nos données ne confirment pas les observations rapportées par Kendon (1967). Il a observé des relations entre les tours de parole et la direction du regard vers le visage ou l'évitement du regard vers le visage. Le regard évité n'a aucune importance dans nos données où presque 100% des fixations sont dirigées vers des

régions sur le visage. Comme notre scénario impose la suite de tours de paroles dans nos expériences, l'apparence des négociations des tours de parole et aussi des effets qui en résultent en tant que comportement spécifique peut être réduite. Nos conditions expérimentales ne sont pas comparables avec celle des expériences de Kendon et ainsi nous ne considérons pas les observations différentes comme contradictoires.

Comparé avec la méthode utilisée par Kendon, notre niveau de résolution de la direction du regard ainsi que la structuration de la conversation en états cognitifs sont beaucoup plus détaillés. Les clignements récurrents des yeux ainsi que la tendance à regarder l'oeil droit durant les états de *pre-phonation* pourraient signifier un comportement comparable à celui rapporté par Kendon. En générale, l'analyse de clignements des yeux montre une relation entre la fréquence d'occurrence et l'état cognitif et le rôle. Les observations suggèrent l'hypothèse qu'une augmentation de l'attention entraîne une réduction de la fréquence des clignements des yeux. Cela pourrait être un signal social d'attention ou bien être du à la nécessité de ne pas gêner la vue et la perception visuelle pendant la réception des informations. Probablement c'est une combinaison des deux. Il y aussi une tendance remarquable que des clignements apparaissent pendant la préparation d'un énoncé, spécialement après l'état cognitive de *reading* dans le rôle *respondent*. Comme cela est accompagné d'un mouvement de la tête vers le haut, les clignements pourraient être un geste de protection (Evinger *et al.* (1994)). Très probablement, c'est un signal dans le contexte de passage de tours de parole, comparable avec le regard évité au début d'un tour de parole que rapporte Kendon. Il a considéré un tel évitement de regard comme un signe d'insister sur la prise de parole par l'inhibition des signaux concurrents de la part de l'interlocuteur.

Les résultats confirment aussi la qualité de notre dispositif expérimental en soi. Selon nos connaissances, c'est le premier qui permet d'enregistrer en détail les deux sujets interagissant dans une conversation en même temps. Les résultats que nous avons obtenus avec un scénario d'interaction plutôt restrictif encouragent l'extension de ces expériences sur plus d'aspects de la conversation. Le scénario devrait être élargi progressivement à une interaction sans restriction qui permette de détailler les relations de bases que nous avons découvertes. Le remplacement des phrases sémantiquement non prédictibles que nous avons utilisées par des phrases normales, une variation de la tâche, la variation du sexe et du statut social des sujets pourraient être des options dans cette perspective. Dès lors, il devrait être possible d'inspecter plus d'états cognitifs que ceux que nous avons discutés dans le scénario actuel.

Indépendamment du progrès dans la mesure du comportement du regard, les modèles développés nécessitent d'être vérifiés. Le modèle que nous proposons devrait être utilisé pour l'animation d'un agent conversationnel animé dans le même scénario que celui utilisé, pour voir comment il influence le comportement des sujets en comparaison des interactions avec notre sujet cible. L'hypothèse que la fréquence des clignements des yeux qui baisse avec une augmentation de l'attention pourrait être testée dans un scénario inversé. Si la fréquence augmente quand l'attention est attendue comme élevée, cela pourrait déranger l'interlocuteur et provoquer des réactions perceptibles.

Le modèle que nous avons développé, basé sur nos observations pendant des interactions médiatisées entre deux personnes, doit être confronté à une interaction face-à-face entre une personne et un agent animé. Des dialogues, tels qu'observés dans les jeux, qui sont orientés vers des tâches proposent un cadre intéressant pour la mesure de l'impact sur la performance de l'augmentation de l'attention. Le jeu de carte que nous avons utilisé dans notre expérience sur la deixis multimodale peut être élargi pour inclure des négociations verbales entre l'agent animé et le sujet, afin de réunir les capacités du regard dans le contexte de l'attention mutuelle et de la deixis dans une seule tâche de collaboration.

Comme but à long terme, le modèle de regard pour l'interaction face-à-face devrait être élargi par la coordination des mouvements de la tête avec les mouvements des yeux et être combiné

avec le modèle du regard dans la perception de scènes (Picot *et al.* (2007)). Ainsi un seul modèle puissant pour l'animation du regard pourrait être développé pour notre tête parlante. La modélisation de dialogue compléterait notre tête parlante pour devenir un vrai agent conversationnel anime opérationnel.

7.6 REFERENCES DU RESUME

- Argyle, M. & Cook, M. (1976). *Gaze and mutual gaze*. London: Cambridge University Press.
- Butterworth, G. (2003). Pointing is the royal road to language for babies. In *Pointing : Where language, culture, and cognition meet* (S. Kita, editor), pp. 9–33. Mahwah, NJ: Lawrence Erlbaum Associates.
- Driver, J., Davis, G., Riccardelli, P., Kidd, P., Maxwell, E. & Baron-Cohen, S. (1999). Shared attention and the social brain : gaze perception triggers automatic visuospatial orienting in adults. *Visual Cognition*, **6**(5), 509-540.
- Evinger, C., Manning, K., Pellegrini, J., Basso, M., Powers, A. & Sibony, P. (1994). Not looking while leaping: the linkage of blinking and saccadic gaze shifts. *Experimental Brain Research*, **100**, 337-344.
- Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, **26**, 22-63.
- Langton, S., Watt, J. & Bruce, V. (2000). Do the eyes have it ? Cues to the direction of social attention. *Trends in Cognitive Sciences*, **4**(2), 50-59.
- Lansing, C.R. & McConkie, G.W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research*, **42**(3), 526-539.
- Lee, S.P., Badler, J.B. & Badler, N. (2002). Eyes alive. *ACM Transaction on Graphics*, **21**(3), 637-644.
- Picot, A., Bailly, G., Elisei, F. & Raidt, S. (2007). Scrutinizing natural scenes: controlling the gaze of an embodied conversational agent. In *International Conference on Intelligent Virtual Agents (IVA)*, vol., pp. 272-282. Paris.
- Surakka, V., Illi, M. & Isokoski, P. (2003). Voluntary eye movements in human-computer interaction. In *The mind's eye: cognitive and applied aspects of eye movement research* (J. Hyönä, R. Radach & H. Deubel, editors), pp. 473-491. Amsterdam, The Netherlands: North-Holland.
- Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S. & Munhall, K.G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, **60**, 926-940.