D 6.3

ANALYSIS OF INTERACTIVE STORYTELLING APPLICATIONS

Project Number	FP7-ICT-231824
Project Title	Integrating Research in Interactive Storytelling (NoE)
Deliverable Number	D6.3
Title of Deliverable	Analysis of Interactive Storytelling Applications
Workpackage No. and Title	WP6 - Interaction and Dialogue
Workpackage Leader	UOA
Deliverable Nature	Report
Dissemination Level	PU
Status	Final
Contractual Delivery Date	31.12.2010
Actual Delivery Date	22.12.2010
Author(s) / Contributor(s)	Birgit Endraß (UOA), Gregor Mehlmann (UOA), Johannes Wagner (UOA), Felix Kistler (UOA), Nikolaus Bee (UOA), Elisabeth André (UOA), Fred Charles (TEES), Jean-Luc Lugrin (TEES), David Pizzi (TEES), Marc Cavazza (TEES)
Number of Pages	34


IRIS

Table of Contents

Abstract

Starting from the conceptual framework presented in Deliverable D6.2, we present a repertoire of interaction modalities and devices with a focus on speech and dialogue which we consider of particular relevance to IS. We then describe two IS applications that have been developed within IRIS that make use of components included in the repertoire.

- EmoEmma, an interactive installation developed within IRIS (Cavazza, Pizzi, Charles, Vogt, & André, 2009) that illustrates an episode from Flaubert's novel "Madame Bovary" and
- The Virtual Beergarden, a new IRIS IS application which was developed to meet the reviewers' recommendations to consider "popular fields such as soap opera".

Furthermore, we report on a number of empirical studies we conducted to evaluate affect-based and gaze-based interaction in EmoEmma and well as interaction in a cave-based version of EmoEmma. In addition, we present our plans to investigate natural language interaction in the Virtual Beergarden using the measurement toolbox developed within Work Package 7.

1. Conceptual Framework for Interactive Storytelling

Interface design guidelines for task-based systems usually do not consider the specific characteristics of IS systems. Unlike task-based systems, IS systems aim at providing interpretational freedom to the users, enhancing their perception of the story, inspiring their curiosity and encouraging their spirit of exploration. Therefore, interaction design guidelines set up for task-based systems should be reconsidered and adjusted to IS systems. Below, we summarize the results of Deliverable D6.2 towards the development of a conceptual framework for interactive storytelling which is based on design dimensions identified in the literature for IS and a detailed analysis of existing IS systems.

1.1 Design Dimensions for Interactive Storytelling

Various attempts have been made to come up with design dimensions to categorize IS systems. (Schäfer, 2004) suggested five categorization criteria for IS: conceptual structure of the story, spatiality and virtuality, degree of collaboration, degree of control, immersion and suspension as categorization criteria. Rowe and colleagues (Rowe, McQuiggan, & Lester, 2007) introduced the notion of narrative presence as a major construct to discuss the difference between educational IS systems and traditional educational systems. They assume that narrative presence is determined by three kinds of factor: narrative-centric factors (consistency, plot coherence, drama and predictability), user-centric factors (affect, motivation, efficacy and control), and interpersonal factors (identification, narrative load, character believability, empathy, and involvement). While their work provides a very promising starting point for framing evaluation questions, however, they do not specifically address interaction design in IS.

In the following, we present five dimensions we use to structure guidelines for interaction design in IS. The design factors were partially drawn from the work described above. However, they are not used for categorizing evaluation questions for IS in general, but focus on interaction design issues.

1.1.1 Basic Usability

As for any interactive system, usability should be considered a basic requirement for IS systems. This helps to avoid flaws in interaction that could easily frustrate the users and lead to the loss of trust in the system. In the area of usability engineering, a number of packages with design guidelines have been proposed, see, for example, (Shneiderman & Plaisant, 2004). While many of these standard design guidelines still apply, some of them need to be revisited for IS in order to take into account the specifics of user experiences with IS systems. For example, to inspire a player's curiosity, a high degree of variation and unpredictability might be desirable in an IS whereas in a traditional user interface design such features should be avoided.

1.1.2 Narrative Engagement

Narrative engagement reflects the degree of how deep the users are interested to drive the story forward and how much they care about what is happening to the characters. Narrative engagement differs from the concept of engagement known in the domain of digital games as follows: while engagement in games can be related to some distinct moments or actions, narrative engagement assumes user involvement in the whole scenario and a feeling of being a part of the story. If, for example, a user is paying a lot of attention to the interaction devices, but at the same time does not actively affect the story plot, there would be a low level of

narrative engagement.

1.1.3 Narrative Immersion

Narrative immersion reflects to what extent users are able to feel in the scene and the characters they impersonate. Narrative immersion goes beyond related concepts, such as immersion in virtual worlds, which focuses on sensory stimuli as a means to make users feel absorbed in the virtual world. Instead narrative immersion refers to the user's absorption in a story and their identification with the characters. A high level of narrative immersion is supported by scene design, character design and interaction design. Narrative immersion distinguishes from narrative engagement: A user may have an interest in how the story develops and actively contribute to plot advancement (narrative engagement) without putting himself into the shoes of a story character (narrative immersion).

1.1.4 Narrative Flow

An IS system should guide users through a meaningful story scenario. As noted by (Rowe, McQuiggan, & Lester, 2007), setting, plot and characters need to be consistent in order not to destroy the believability of the story. Consistency also refers to the use of interaction devices and modalities which should match the story. For example, using a mobile phone to communicate with a character from the 19th century would be in conflict with the user's conception of the story. In addition, events and actions should follow a logical and causal order, see (Rowe, McQuiggan, & Lester, 2007). That is the user should understand how a story evolves.

1.1.5 Conversational Flow

Conversational flow indicates to what extent the users are able to converse with the characters in a continuous and natural manner. Due to deficiencies of current technology to process natural language input, effective strategies need to be found in order to support a consistent and coherent conversational flow. Based on an evaluation of the desktop-based version of Façade, (Mehta, Dow, Mateas, & MacIntyre, 2007) come up with a number of guidelines and recommendations for dialogue design in IS, such as avoiding shallow confirmations of user input and supporting the user's abilities to make sense of recognition flaws. Another important issue is the timing of user input and system output.

1.2 Characterization of Interaction Modalities and Devices for Interactive Storytelling

In this section, we briefly review typical interaction devices and modalities used in IS. A more comprehensive list may be found in the IRIS repository on our web page.

1.2.1 Text-Based Input

Free text input has certain advantages and disadvantages. On the one hand, free input encourages creativity, enabling users to develop ideas more broadly. On the other hand, free input yields many sources of recognition errors. First, typed sentences may contain spelling errors. Second, the interpretation mechanism may be unable to find a suitable reaction to successfully parsed input. Keyword-based text input is much less error prone. Even more robust performance can be reached with menu-based input. However, such interaction styles

are easily received as boring, less intuitive and restrictive.

1.2.2 Speech-Based Input

Free speech input can be seen as the most natural way for a dialogue, mirroring the way humans interact in real life. However, interpretation errors are even more probable for speech input than for text-based systems.

As for the typed variants, keyword-based speech input is less error prone than free input. If the system is trained accurately and the restricted speech input fits the story scenario, keyword spotting may support harmonic user interaction.

1.2.3 Gesture-Based Input

Gesture-based interaction usually implies a limited set of gesture-keywords that trigger certain actions. If the system is properly trained and the gesture recognition rate is sufficiently high, this kind of interaction can positively contribute to user engagement. Since gestures imply bodily activity, user involvement into the story is likely to be increased. Moreover, bodily activity may enhance the user's feeling of playing a specific role.

1.2.4 Gaze-Based Input

Eye gaze plays an important role in face to face conversations. Eye gaze can give feedback to an interlocutor and regulate and synchronize the flow of a conversation, see (Argyle & Cook, 1976), (Kendon, 1967) and (Kleinke, 1986). In IS, eye gaze tracking may be used to realize more natural agents that are aware of the user and notice where he or she is looking. For example, when users start to stare at the agent, which is often the case in systems where users interact with virtual characters, see, for example, (Rehm & André, 2005), the virtual agent should naturally avert his gaze as humans would do in social interactions.

1.2.5 Direct Manipulation

Direct interaction is also known to enrich user involvement into the story. Interaction by means of tangible objects or graspable devices that fit the story scenario enhances user experience and makes them feel the story to be "closer". This approach is especially valuable for IS systems for children: young users appreciate graspable toy-like objects. This kind of interaction is easy to understand: control over the digital story is achieved via objects similar to those known from real life. No technical bridges need to be built in order to understand abstract concepts as it might be the case for keyboard, mouse or other auxiliary interaction devices.

2. Technical Interaction Framework Including a Repertoire of Interaction Devices and Modalities

In this section, we introduce a technical framework that contains components to analyze the user's input to an interactive story telling system with focus on natural language processing and dialogue control. The selection of the components is based on the conceptual analysis provided in Section 1.

2.1 Dialogue Manager

The SceneMaker tool facilitates the dialog management and the creation of interactive performances of virtual characters. It divides the authoring task into the creation of dialog content and the modeling of the narrative structure of an interactive performance.

Dialog content is organized in a set of scenes that are specified in a multimodal scene script which resembles a movie script with dialog utterances and stage directions for controlling gestures, postures and facial expressions.

The narrative structure of an interactive performance and the interactive behavior of the virtual characters is controlled by a scene flow - a state chart specifying the logic organization and temporal order in which scenes are played and commands are executed. Our improved version of scene flows provides concepts for hierarchy, concurrency, variable scoping, multiple interaction policies and a runtime history.

Scene flows and scene scripts are created using our graphical authoring tool, which supports authors with drag 'n' drop facilities to draw the scene flow by creating nodes and edges. In addition, it allows to annotate nodes and edges with statements and expressions from a simple scripting language and includes a scene script editor with syntax analysis and syntax highlighting features as well as an integrated gesture lexicon.

The first approach for the execution of Scene flows relied on a compiler that translated the model into an executable program (Gebhard, Kipp, Klesen, & Rist, 2003) while the current version of SceneMaker relies on an interpreter for the execution of scene flows. This allows the real-time extension and modification of the model and the direct observation of the effects without the need for an intermediate translation step. The interpreter also allows the real-time visualization of a scene flow's execution and active scenes within the graphical user interface in order to test, simulate and debug a model in order to control the modeling progress and to verify the correctness of the model.

2.2 Language Understanding

In order to integrate natural language understanding into interactive story telling scenarios, we use the Semantic Parser SPIN. This parser is especially suitable as it enables the analysis of free word order languages and as it copes well with faulty or incomplete text input. We think this is of special importance when developing an interactive storytelling system, as the user's input is mainly spontaneous in this domain and not necessarily well formulated. In addition, we integrated a spell checker in order to assure that the system understands utterances that contain typos.

The goal of using a parser in our component-based interactive storytelling system, is to analyze the user's input and to parse it into abstract dialog utterances, which are sent to the

behavior modeling tool for further processing. From our conceptual framework we learned that we should provide clear interaction prompts. Thus, the characters are modeled in a way that they are asking the user for suggestions at certain points in the story. In that manner, on the one hand the user knows what kind of input is required. On the other hand, the parsing of these sentences is a lot easier, since the domain is limited.

In order to achieve correct parsing of the user's input, rules that match the dialog utterances need to be defined as well as a lexicon that holds all possible words, syntactic and semantic categories.

2.3 Full Body Interaction

For providing full body interaction we are using Microsoft Kinect. Kinect is an additional peripheral (precisely a sensor) for the Microsoft Xbox 360 (a video game console). It provides an intuitive interaction similar to the Nintendo Wii Remote (a controller with accelerometers and infrared detection for another video game console, the Nintendo Wii), but without any controller. The user can interact with his whole body or object in his hands and additionally with spoken commands.

2.3.1 Hardware

The Kinect consists of a RGB camera, a multi-array microphone, a tilt motor, a status LED, an infrared (IR) camera and an IR emitter.

- The IR camera and the IR emitter are used for providing a depth image that is created according to the structured light principle. For this matter, the IR emitter projects the scene with an irregular pattern of IR dots and the IR camera reconstructs a depth image by recognizing the distortion in this pattern.
- The status LED changes its color according to Kinect's connection status.
- The tilt motor is used to adjust the viewing angle of the sensor.
- The multi-array microphone uses ambient noise suppression, can recognize spoken commands and additionally localize their acoustic source.
- The RGB camera mainly acts like a normal webcam, but its image can be registered to the depth image for combining the depth and color information of the scene.

2.3.2 Open Source Drivers

The first open source drivers named Freenect were released on the 10th November 2010 by Hector Martin and later renamed to Open Kinect. They are further developed by the community, supporting to connect the Kinect over USB to all major platforms (Windows, Mac, Linux) and they can be used with a wide range of programming languages (e.g.: C/C++, C#, Java, Python). They currently provide the RGB, IR and depth image of the Kinect.

On the beginning of December 2010 PrimeSense, the manufacturer of the 3D sensor technique of the Kinect, released, in cooperation with some other organizations, their own open source drivers, a framework called OpenNI (Open Natural Interaction) and a middle ware called NITE (Natural InTEraction). The three of them also provide the RGB, IR and depth image, but additionally the registration of the depth and RGB image as well as user body tracking and gesture recognition. They support Windows and Linux as platform and C/C++ as programming language.

2.4 Gaze-Based Interaction

In order to enable natural gaze-based behaviors in IS, we integrate eye gaze tracking into an agent's interactive eye gaze model, see (Bee, et al., 2010) for more details.

2.4.1 Eye Gaze Tracking

Many systems investigating interactive models of visual attention make use of head trackers, see (Nakano, Reinstein, Stocky, & Cassell, 2003) or (Sidner, Kidd, Lee, & Lesh, 2004). They are able to roughly assess in which direction the user is looking, but do not have more detailed information on the user's eye gaze direction. In our work, we make use of the SMI iView X RED eye tracker. It operates with a sampling rate of 50 Hz and the tracking accuracy is less than 0.5°. The distance between the eye tracker and the user should be about 60 - 80 cm. The advantages of an unobtrusive, contact-less eye tracker include that users do not have to wear a sometimes bulky apparatus and thus are not steadily reminded that their gaze is tracked. Further, the SMI iView X RED eye tracker allows head movements horizontally and vertically up to 20 cm in each direction.

To find fixations, we use the I-DT algorithm by (Salvucci & Goldberg, 2000). According to I-DT, a fixation is detected when the eye coordinates of a frame lie within the distribution disp. For each frame disp is calculated with the following formula: $disp = (max_x - min_x)+(max_y - min_y)$ where min_x, max_x, min_y and max_y are the minimum and maximum coordinate values of all points inside the frame. If disp is beyond a certain threshold the current frame is detected as the beginning of a fixation and then expanded by following points until the threshold is exceeded. This marks the end of a fixation. The samples in the final window are averaged to a single fixation point.

2.4.2 Eye Gaze Model

A number of studies that investigate the role of eye gaze in human-agent communication provide evidence that natural eye gaze behaviors of an agent that is informed by studies of human-human conversation are not only more positively perceived, but elicit more natural responses in human users (see, for example, (Colburn, Cohen, & Drucker, 2000), (Garau, Slater, Bee, & Sasse, 2001), (Lee, Badler, & Badler, 2002) or (Vinayagamoorthy, Garau, Steed, & Slater, 2004)). In our work, we start from the gaze model developed by (Fukayama, Ohno, Mukawa, Sawaki, & Hagita, 2002) which allows us to specify a number of gaze parameters that influence the impression a character conveys. Their model includes two states: looking at the user and averting the gaze from the user. Three parameters define how often, how long (500 to 2000 ms) and where the virtual agent looks. The gaze targets consist of a set of random points from either all over the scene, above, below or close to the user. The probabilities of changing from one state to the other or staying in the same state depend on the amount and the mean duration of the gaze parameters. (Fukayama, Ohno, Mukawa, Sawaki & Hagita, 2002) rated the impression particular gaze patterns conveyed that were produced by modifying the gaze parameters.

They found that a medium amount of gaze and a mean duration between 500 to 1000 ms conveys a friendly gaze behavior. The orientation of the gaze direction did not play a decisive role in distinguishing between friendly and dominant gaze behavior, except a downward gaze was considered as less dominant. Fukayama and colleagues evaluated their gaze behavior model by only displaying eyes to the users. Thus, we evaluated their model with a full virtual character that moves his head as well as his eyes. Basically, we followed their settings, but distinguished whether the agent is speaking or listening.

Our gaze model was extended with further parameters as our virtual agent is capable of reacting to the user's current gaze using an eye tracker. The maximal and minimal duration of mutual gaze can now be set as well. Furthermore, we may indicate the maximal duration the

virtual agent gazing around.

2.5 Emotional Speech Recognition

Emotional speech extracted from the voice is analyzed by EmoVoice (Vogt, André & Bee, 2008). Real-time recognition of vocal emotions is a three-step process. First, the acoustic input signal coming continuously from the microphone is segmented into chunks by Voice Activity Detection (VAD), which segments the signal into speech frames with no pauses within longer than about 0.5 seconds. Next, from this speech frame, a number of features relevant to affect are extracted. The features are based on pitch, energy, Mel Frequency Cepstral Coeffcients (MFCC), the frequency spectrum, the harmonics-to-noise ratio, duration and pauses. The actual feature vector is then obtained by calculating statistics (mean, maximum, minimum, etc.) over the speech frame ending up with around 1300 features. A full account of the feature extraction strategy can be found in (Vogt, André & Bee, 2008).

In the last step, the feature vector is classied into an affective state. Integrated classifiers are Support Vector Machines (SVM) and Naïve Bayes (NB), while the latter is used more often because it is faster and thus responds better to real-time demands. The NB classifier is very fast, even for high-dimensional feature vectors and therefore especially suitable for real-time processing. However, it has slightly lower classification rates than the SVM classifier which is a very common algorithm used in offline emotion recognition.

In combination with feature selection and thereby a reduction of the number of features to less than 100, SVM is also feasible in real-time.

3. Applications Developed with the Technical Interaction Framework

In the following, we describe how components of the technical framework have been used to realize the interaction in two interactive storytelling applications.

- **EmoEmma**

  The interactive installation developed within IRIS (Cavazza, Pizzi, Charles, Vogt, & André, 2009) illustrates an episode from Flaubert's novel "Madame Bovary". The main character, Emma Bovary, is bored by her life at the country side with her husband, a local doctor. One day a rich landlord, Rodolphe, visits her. Emma feels empathy towards Rodolphe and is ready to go for a story with him. The IS system represents the scene of Rodolphe's visit. Two persons participate in the scene: the virtual characters of Emma and Rodolphe, see Figure 1. The dialogue between them is supposed to decide the future development of their relationship.

- **The Virtual Beergarden**

  The virtual Beergarden (Endraß, Boegler, Bee, & André, 2009) is a new IRIS IS application which was developed to meet the reviewers' recommendations to consider "popular fields such as soap opera". It features a virtual dinner party with virtual characters. The user impersonates one of the characters in the virtual Beergarden, can move around from one group to the other, engage in a conversation with the characters and thus contribute to the ongoing drama. The underlying narrative structure of the Virtual Beergarden is created with support from a professional script writer who was hired following the reviewers' recommendations. Figure 2 shows a screenshot of the scenario.

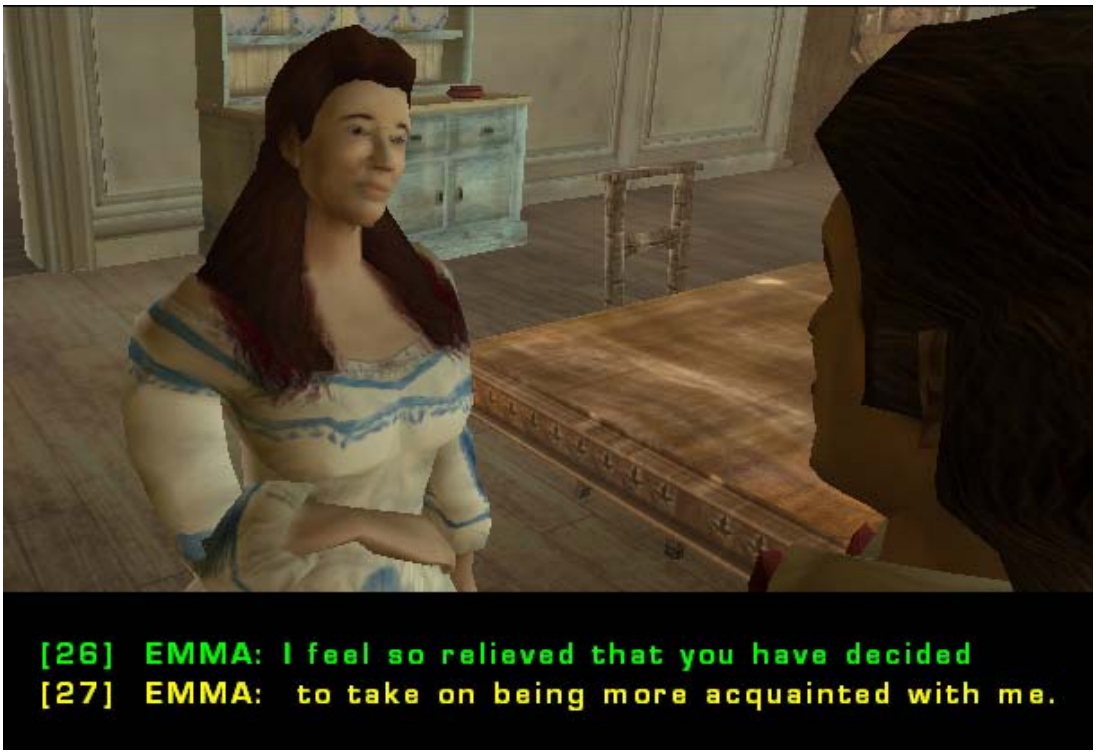

Figure 1: Screen shot of EmoEmma

Figure 2: Screenshot of the virtual Beergarden

3.1 Interaction in EmoEmma

During a typical demonstration session, dialogues initiated by Emma provide opportunities for user interaction, since they tend to prompt the user to reply. Because user utterances are interpreted in terms of their emotional impact, their consequences are seamlessly deferred until the appropriate evolution of Emma's emotional state through the narrative. For instance, when Emma begins Rodolphe to take her away, the user can respond more or less enthusiastically and this may bring consequences when at later stages Emma is finally making her choice between Rodolphe and her husband. On the other hand, the user's response to Emma's love declaration is more likely to trigger an immediate reaction. The system is also able to make sense of the user's silence, but only interprets it meaningfully for some crucial questions or requests from Emma. The IS system is displayed on a screen (desktop screen or large immersive vertical screen). The character of Emma is located in the centre; she talks to Rodolphe by means of speech. The user provides the phrases from Rodolphe's side by means of a natural speech input with a microphone.

Multimodal interaction in EmoEmma follows the conceptual framework in Section 1. Table 1 gives examples of strategies we applied to satisfy the requirements of the framework dimensions.

Basic Usability	Narrative Engagement	Narrative Immersion	Story Flow	Conversational Flow
Subtitles	Emotional involvement	1st person perspective; means to express emotions	Emotional Interaction matches the style of a love scenario	Interaction leaves space for interpretation

Table 1: EmoEmma interaction strategies

3.1.1 Affective Interaction in EmoEmma

Speakers express different emotions depending on the application they interact with. EmoVoice integrates an easy-to-use interface for recording and training an emotional speech corpus, which is meant to increase accuracy in application-dependent contexts. The method used for emotion elicitation was inspired by the Velten mood induction technique (Velten, 1968) as used in (Vogt, André & Bee, 2008) where subjects have to read out loud a set of emotional sentences that should set them into the desired emotional state. The system comes with a list of such sentences, which we have completed with actual excerpts from Madame Bovary's dialogues.

For EmoEmma, we have concentrated on a small set of five categories (each corresponding to combinations of valence and arousal): *NegativeActive*, *NegativePassive*, *Neutral*, *PositiveActive* and *PositivePassive*. The rationale for such a reduced set of emotional inputs has been that these categories will be further interpreted, taking into account the context in which they are recognised. However, developers making use of the system are encouraged to change sentences according to their own emotional experiences: For a good speaker dependent system, about 40 sentences per emotion are usually sufficient.

We have initially trained EmoVoice with three subjects using various test sentences, some of which extracted from the actual dialogues of *Madame Bovary*, with an average of 40 sentences per category. Overall we have achieved a recognition score of 66% for those five categories, obtained with speakers outside those having contributed to the system training. This score is consistent (and probably on the upper end) with those previously reported for the EmoVoice system (Vogt, André, & Bee, 2008).

User utterances are interpreted contextually as a function of the relation between the EmoVoice category and the character's expectation. The affective response to the user's reply is amplified by Emma's current emotional status: for instance, a lukewarm attitude from Rodolphe would upset Emma all the more that her expectations run higher at any given stage. For instance, in case of high expectations from Emma, *NegativePassive* and *NegativeActive* utterances will be interpreted as feelings of disappointment, rated in levels of intensity determined by the *Active*/*Passive* component. This is also a mechanism to incorporate the dynamics of the relationship, as expectations would vary according to the status and progression of the characters' relationship throughout the narrative. A similar affective response would have dramatically different effects at various stages of the unfolding narrative.

3.1.2 Integrating Gaze-Based Interaction into EmoEmma

An eye tracker was connected to enable the interactive gaze model described earlier to respond to user's current gaze (i.e. looking into the virtual character's eyes or not).

For the character, we implemented the following interactive eye gaze behavior. The character looks for about 2 s (between 1 and 3 s) at the user before she averts her gaze again for about 4 s (between 2 and 6 s). Whenever the user is looking at Emma, she tries to establish mutual gaze and to hold it for about 1 s (between 0.75 and 1.25 s). The duration of gaze to and away from the user is slightly adapted depending on whether the agent is talking or listening; taking into account the fact that people look more at the interlocutor when listening than when talking, see (Argyle & Cook, 1976).

Figure 3 shows a typical set-up of gaze-based interaction with EmoEmma. The user is placed in front of a table on which the eye tracker is placed. The projection surface sizes 120 × 90 cm, which displays the virtual agent in life-size. To offer an enriched scene where the user has the choice to look away from the virtual agent, Emma is placed in the dining room of her house, which includes chairs and tables. Before the interaction starts, the eye tracker is calibrated, which takes less than 2 minutes.

Figure 3: Setting for gaze-based EmoEmma

3.2 Multimodal Interaction in the Virtual Beergarden

The ultimate goal in the virtual Beergarden is to enable natural conversational behaviours using a combination of speech- and gesture-based input. Currently, two versions of the Virtual Beergarden have been realized:

- A text-based version in which the user may navigate using the mouse and freely interact with any of the characters by typing utterances
- A motion-based version in which the user may physically approach the characters and communicate with them using hand gestures

In both versions, the interactive behaviour is controlled using SceneMaker.

Multimodal interaction in the virtual Beergarden is based on the conceptual framework presented in Section 1. Table 2 gives examples of strategies we applied to meet the

requirements of the framework dimensions.

Basic Usability	Narrative Engagement	Narrative Immersion	Story Flow	Conversational Flow
Human-like gestures and spatial behaviours	Characters' response to full body motion and language input → effectance	1st person perspective; full body motion	Interaction matches the style of small talk	Clear interaction prompts; space for interpretation

Table 2: Interaction strategies in the virtual Beergarden

3.2.1 Modelling Interaction with SceneMaker

To find the prototyping of the sceneflow for our first version of the SOAP scenario, we first identified the objects within the Beergarden setting and the different processes that are proceeding within the setting before we linked the running processes to our objects.

This approach is the general approach when modeling those scenarios with concurrent and hierarchical statecharts whereby an author can take on a more object oriented view on the setting or a more process oriented view, however the result of her deliberations is usually some mixture of both. In the first place, it might be useful to identify objects and implement the behavior, the inner states and attributes of such an object in a single process. However, in order to reduce the modeling effort and the model complexity, it might also be useful to group objects into clusters each of which is actually performing a single process. On the other hand, it might be necessary to refine the performance of objects into several concurrently proceeding processes, modeling behavioral aspects of that object. For our first version we finally identified four main processes proceeding in the virtual Beergarden scenario at the beginning of the game which are explained in the following. The automata modeling those processes at the highest level of the node hierarchy of the SOAP sceneflow are shown in Figure 4.

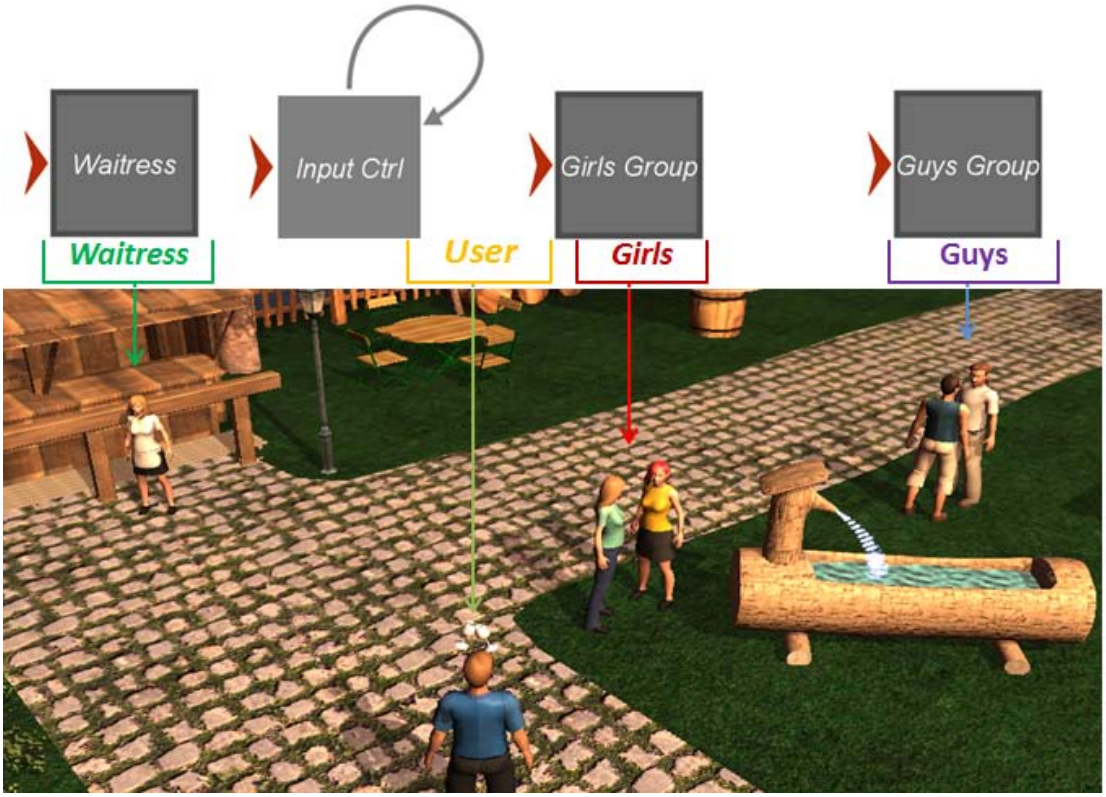

Figure 4: The highest level of the SOAP sceneflow's node hierarchy

a. The User Agent

The user avatar is controlled by the user and can start a conversation with the other agents in the virtual Beergarden. The automaton is listening to and evaluating the user's discourse acts dependent of the context of different focus group and the progress of the current dialog. A simplified implementation of the user control automaton is shown in Figure 5.

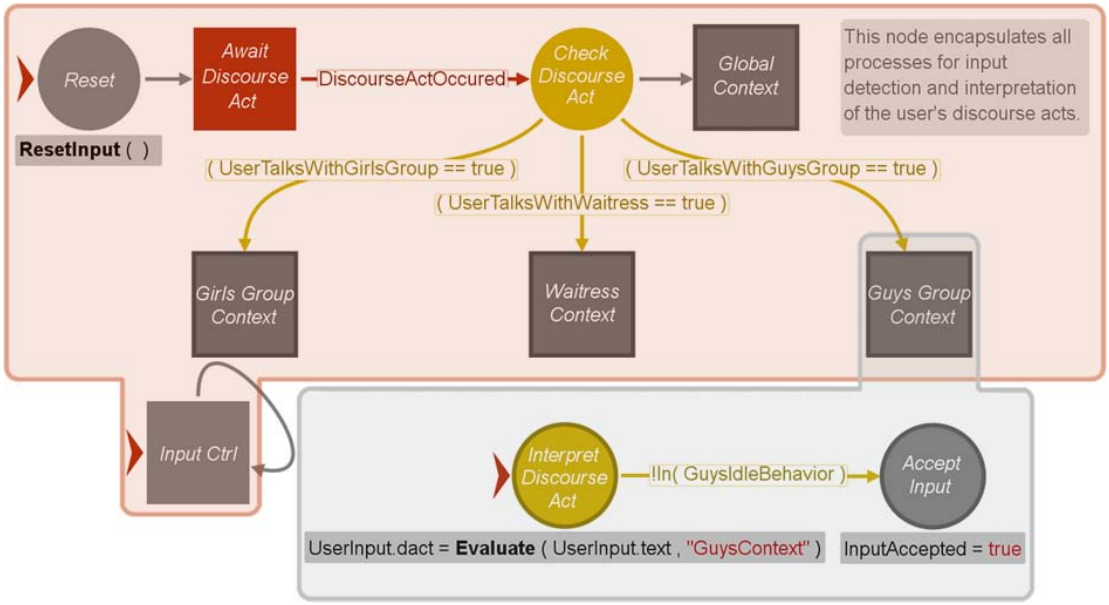

Figure 5: Simplified version of the automaton evaluating the user's discourse acts dependent on the focus group and the progress of the current dialog.

b. The Waitress Agent

The waitress agent's behavior is modeled in a separate automaton, named Waitress. The waitress is periodically going through the virtual Beergarden asking the guests for orders until she is involved in a conversation with either the user or one of the guys. In case of a conversation with the user, the process modeling the waitress needs to be properly synchronized with the user agent, which means the interactions of the user. The case of the conversation with one of the guys represents a change in the focus group constellation usually also initiating the transition from one story scene to another one.

c. The Girls Group Agent

The Girls Group Agents are clustered into a group which is modeled in a single automaton. The Girls Group automaton is shown in Figure 6 and is modeling the idle behavior of the girls' group agents as well as the conversation between the girls and the conversation of the girls with the user's agent, dependent of the user's current dialog focus group.

In case of a conversation with the user, the process modeling of the girls group needs to be properly synchronized with the user agen. The synchronization is realized by reacting to the user's different discourse acts during the conversation. After a discourse act has been detected, the conversation either proceeds with the transition to another dialog topic or with reentering the last dialog topic.

The reopening of a dialog is modeled using the history mechanism of SceneMaker as shown in Figure 6 and Figure 7.

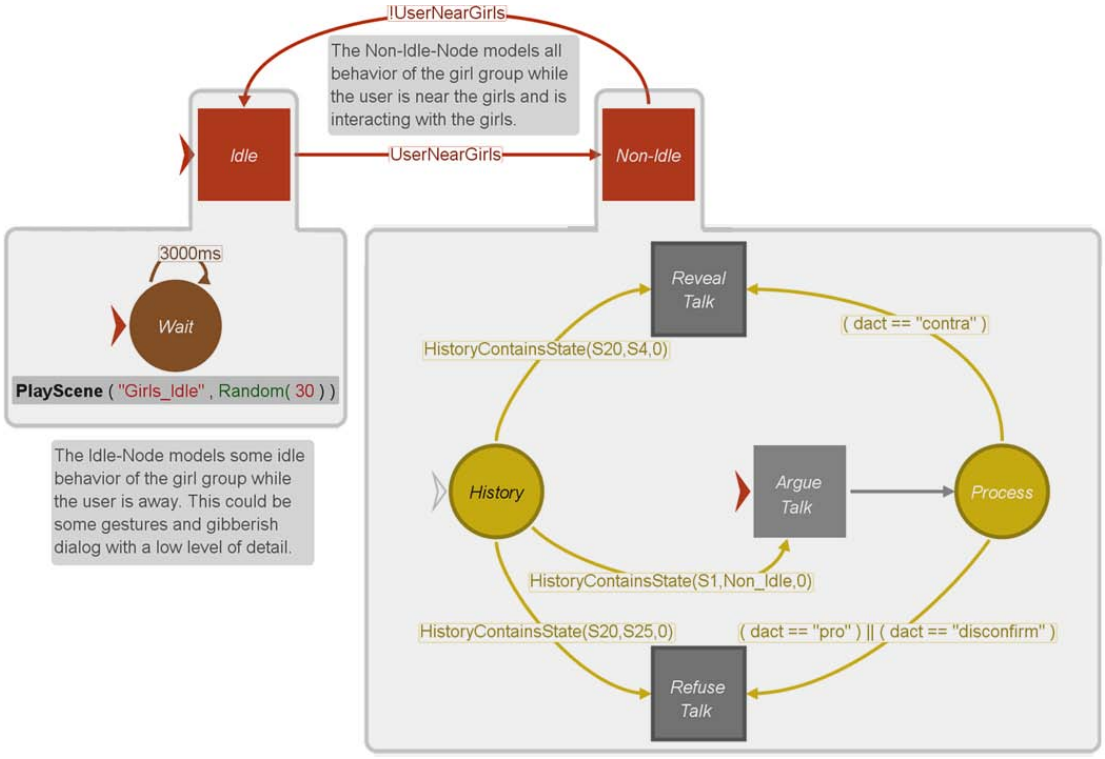

Figure 6: The behaviour of the girls dependent on the user's discourse acts

Figure 7 shows a more detailed look on an automaton modeling the discussion of a single dialog topic between the user and the girls group. It can be seen that the girls' reaction to a

certain discourse act, as for example confirmation, interruption or greeting is realized by playing a scene from a scene group. After the play back of that scene the dialog is either reentered or the current dialog topic is terminated.

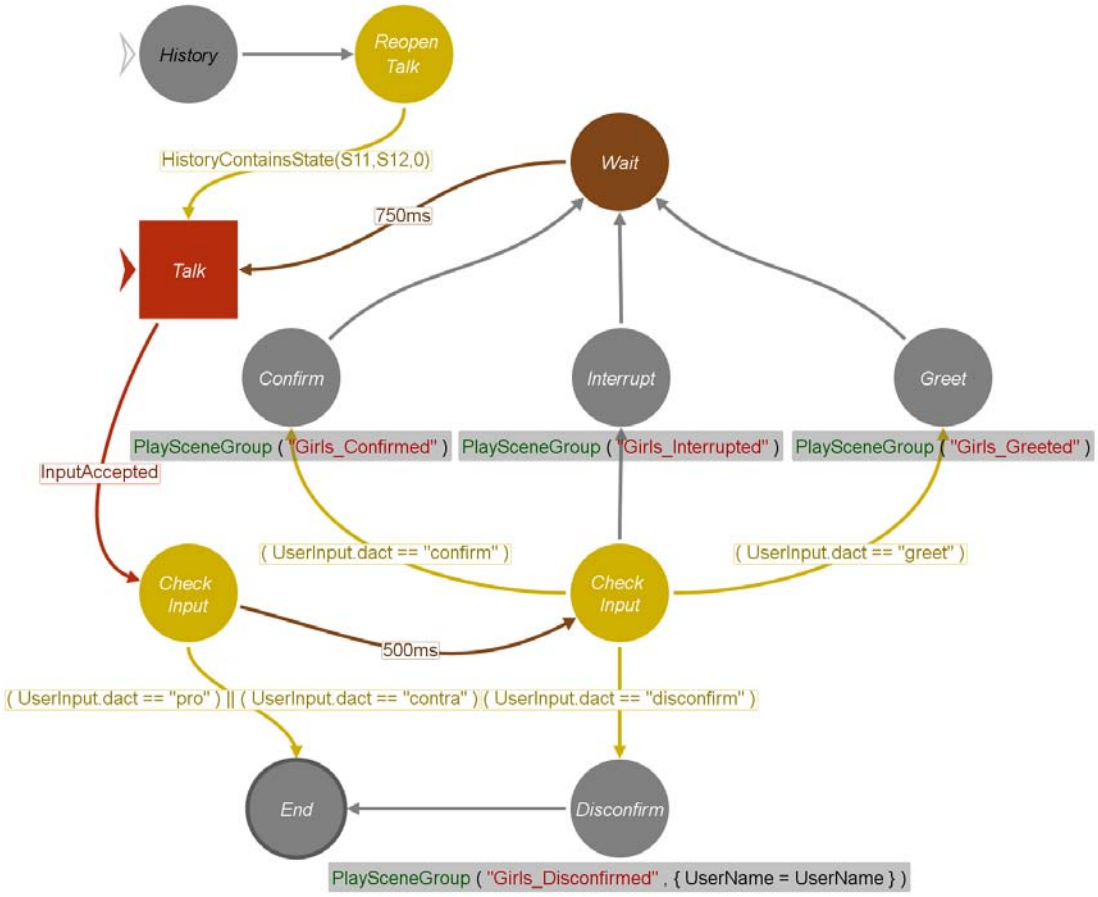

Figure 7: Simplified version of the automaton specifying the girls' behaviour in the conversation

d. The Guys Group Agent

The Guys Group Agents are clustered into a group which is modeled in a single automaton as well. The structure of the Guys Group automaton is very similar to the Girls Group automaton and is not further explained here.

In the further course of the game, there are rising dynamically changing constellations of the conversational agent groups. Agents are leaving the one group and entering another group. These changes can then be modeled in further automata. Since those constellation changes usually represent a transition from one story scene to another or at least the switch from one plot sequence to another sequence, the author can provide automata for each of the possible scene constellations, thus modeling possible subsequent scenes in separate automata, that could be executed by a drama management module.

3.2.2 Language Understanding in the Virtual Beergarden

To enable the user to engage in free style natural language dialogue with the virtual Beergarden characters, we use the semantic parser SPIN.

Figure 8 exemplifies a simple rule for our scenario using the parser. The rule states that when the user's typed text input contains one of the words "how", "do" or "what" in correlation with

the word "you" and any word belonging to the semantic category location, the abstract speech utterance "ask-location" is triggered and send to the behavioral model of the target virtual character. For this rule, a semantic category location has to be defined as well. In the example in Figure 8 this category contains the words "location", "place", "beergarden", "here" and "party". Thus, user input such as "How do you like it here?" is matched to the dialog utterance "ask-location". Please note, that faulty input such as "porty" instead of "party" is corrected according to the words defined in the lexicon.

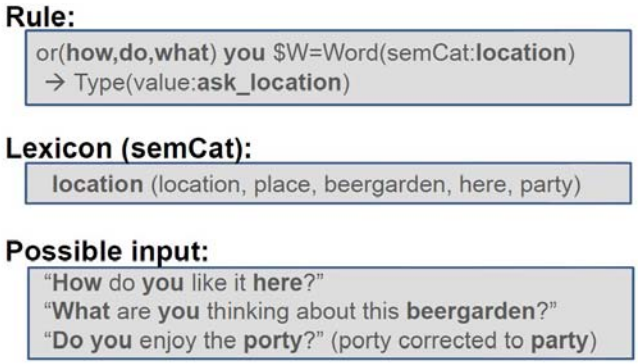

Figure 8: Example rule with corresponding user input as defined by the semantic parser SPIN

In addition, word stems and other pre-processes can be defined. As shown in Figure 9, conjunctions can be summarized in one word stem or different formulizations can be preprocessed to match the common meaning. With this, the system is able to understand different utterances with the same meaning, such as "I like this place" and "I am liking this place".

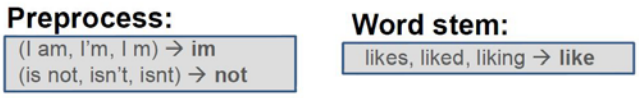

Figure 9: Example of pre-processes and word stems in SPIN

3.2.3 Kinect Interaction in the Virtual Beergarden

Kinect-based interaction in the Virtual Beergarden (see Figure 10) was realized with the Freenect Open Source driver under Windows and with C++ as programming language. We integrated the Kinect drivers into a toolbox for social signal processing developed by Augsburg University, which analyzes the depth image, extracts low level features, such as the distance of the user to the camera as well as some higher level features, such as gestures performed by the user with his hands and sends them over a socket connection to the Virtual Beergarden.

Using the Kinect, we allow users to freely navigate through the Beergarden (without using a mouse or other interaction devices) and join or leave groups of other agents. The agents respond to the human user by appropriate spatial behaviors. One determinant of such spatial behavior is the acceptable distance between communication partners. While the human user approaches the agents, they try to keep an appropriate social distance. Furthermore, the agent signals human users by their spatial behavior whether they welcome them as new group members or not. This behavior is realized by interpreting the distance of the user to the Kinect sensor.

In addition, the user may perform various gestures with his hands which are recognized with the $1 algorithm. The agents can react to the gestures. For example, they greet and wave back to the user if he waves with his hand. As the Freenect driver only supports the raw depth image, we used some simple techniques like a threshold and some simple segmentation over the depth pixels, to extract the mentioned features.

We are also planning to integrate the newer driver, the framework and the middle ware released by PrimeSense, which provide full user body tracking and better gesture recognition.

Figure 10: Kinect Setting

4. Evaluation of Interaction

In this section, we summarize the results of various studies we conducted to evaluate the interaction devices and modalities used in EmoEmma and the Virtual Beergarden. The purpose of the studies was to investigate which influences the interaction styles had on human participants.

4.1 Affective Interaction with EmoEmma

In order to evaluate EmoEmma, we prepared an experiment where users interacted with the character using the emotional tone in their voice. The objective of the evaluation was to find out which effect a pure emotional interaction had on the users' perception of the experience and the attitude towards the character.

4.1.1 Subjects and Setting

We recruited 14 subjects for the experiment. The setting consisted in the interactive narrative being displayed on a 30'' screen, with a high-quality microphone positioned in front of it. Subjects were first asked to read aloud several excerpts from the original novel's dialogues, in order to test optimal acoustic signal strength for emotional speech recognition. They were given instructions describing the narrative, the part they were supposed to play impersonating Rodolphe, and the fact that Emma would react to the emotional content of their responses (however, they were not given any detail on the actual techniques underlying the system, such as the fact that it did not recognise word meaning). The IS system was then started, generating real-time 3D animations, with a voice over giving the background for the early stages of the narrative at which no interaction was allowed. The user had no control over navigation of his character, and was presented in third person mode, with Rodolphe as his avatar. An automatic camera system (part of the visualisation engine) was centred on Emma *and would follow her on stage.*

As the IS system starts, generating the first encounters between Emma and Rodolphe, the user can either address Emma spontaneously or respond to one of her questions or declarations which are enacted through a corresponding animation with text-to-speech voice synthesis. After each utterance from Emma, the user has the choice of responding her with various level of enthusiasm, empathy or disapprobation, or not to respond, which in some cases will also give raise to an interpretation, based on the level of Emma's expectation. The user can experience the subsequent unfolding of the interactive narrative, whether he continues to interact or not: his replies may show immediate or deferred effects or no effects at all. Multiple interactions are allowed throughout the scene, as Emma repeatedly addresses Rodolphe as part of her role/plan. At no stage does the user receive any indication of the emotional category perceived from his utterance, and his only feedback is via the interactive narrative *itself.*

4.1.2 Results

All 14 subjects successfully completed the experiment, which resulted in each case in a complete session, generating an interactive narrative until its normal ending. The average duration of the interactive narrative was 2.9 minutes (with extremes varying from 2 to 6 minutes) and ended up with either Emma leaving the stage in despair ("negative" ending), or engaging with Rodolphe ("positive" ending, which actually occurs in the original novel). Subjects were not instructed to favour a particular outcome, nor were they described any given outcome as normal: as a result, their interventions were balanced in nature, leading to

an almost equal split between each possible ending (57% positive versus 43% negative). The actual sequence of narrative events was of much greater variability and its constituency depended on the nature and number of user interventions. Longer stories emerged as the user gave successive contradicting messages, which lead Emma through opposite feelings, provided none is so extreme as to accelerate the ending. In a similar fashion, high intensity emotional categories (active), regardless of their valence tended to lead more quickly to the story ending. An alternative explanation would correspond to users trying to correct the impact of EmoVoice recognition errors, which they perceive through inappropriate responses from the Emma character, by repeating a similar type of utterance to the one they see as having been unsuccessful. However, because of the relative robustness of the system, and the rather unconstrained nature of the experiments, these contradicting messages could correspond to exploratory behaviour by the subjects.

One explanation for the overall system robustness despite a 66% emotion recognition score can be found in the actual type of recognition errors. A study of system logs during user experiments showed that the most severe errors, in which opposite valence categories such as *NegativeActive* are recognised instead of the reference *PositiveActive*, only occur in about 5% of utterances. Most of the errors do not affect valence and, notwithstanding the value of expectations at the point at which they occur, tend to produce similar results in terms of narrative impact, or more often have no impact at all, and the dynamics of *the story offers 'second chances' for correcting this seamlessly.*

These evaluations do not aim at measuring the intrinsic aesthetic quality of the novel, but tend to validate the overall concept, and assess user engagement with the system. The average number of interactions (user utterances) during a session was 7.4 ± 5, and there was no clear correlation between interactive narrative duration and the number of interventions (which can be accounted for by the redundancy of some interventions). The average length of user utterances was 7.5 words with a significant proportion of utterances exceeding 10 words, again suggesting that the users were comfortable interacting with the system.

Questionnaires after the interaction revealed that the subjects responded very positively to the installation and perceived EmoEmma as a believable character that responded appropriately to what they were saying. E. g. on a scale from 0 to 5 they rated 3.6 on average that Emma understood what they were saying. When interpreting this result, we should keep in mind that EmoEmma did not analyze the semantics of the user's utterances, but solely aimed at recognizing the user's emotions from the acoustics of speech. Thus, the result of the user study can be taken as evidence that EmoVoice was effective since it was the only mode of interaction.

A detailed description of this study is given in (Cavazza, Pizzi, Charles, Vogt & André, 2009).

4.2 Affective Interaction with Cave EmoEmma

Further user experiments were carried out with a fully immersive CAVE-like EmoEmma, which explores two interactivity paradigms for user involvement (Actor and Ghost). To this end, all components of EmoEmma were integrated in an immersive setting (a 4-screens CAVE-like stereoscopic display).

We have defined two modes of interaction: in the Actor mode, the user becomes a member of the cast and is expected to play a role, while in the Ghost mode she can navigate freely as an invisible character, still able to interact with the world and the virtual actors. Our objective was to assess whether users are actually able to interact successfully and in a relevant fashion, with such a complex environment.

4.2.1 Participants and Setting

Thirty-eight users participated in this experiment, including twenty male and eighteen females with an average age of 30.6 years with a range between 19 and 57 years. The average

duration of a session was 45 minutes.

The experiment consisted of five main parts: i) Story introduction (~10 minutes); ii) Practice session (~10 minutes); iii) Experiment sessions A (~6 minutes); iv) Experiment sessions B (~6 minutes); v) Completing questionnaires (~15 minutes). During the first part, participants were asked to watch a desktop version of the interactive narrative which would demonstrate the type of interaction between characters (both verbal and mediated by objects). Then the two interaction paradigms, Actor mode and Ghost mode, were presented as two separate scenarios. In order to minimise any learning curve effects, the order of these scenarios was randomly chosen for each candidate. However, during this briefing great care was taken not to disclose how user actions may influence the interactive narrative. Subjects were offered a practice session to get acquainted with navigation and interaction in a virtual world.

The third part represents the core of the experiment, where users experience immersive interactive storytelling. As discussed in the Results section, for each session the user's interactions, navigation and vision were constantly logged. Finally, participants were requested to complete two questionnaires immediately after their last session. The first one is the Simulator Sickness Questionnaire (SSQ) developed by (Kennedy, Lane, Berbaum, & Lilienthal, 1993), while the other one is the ITC-SOPI Presence questionnaire proposed by (Lessiter, Freeman, Keogh, & Davidoff, 2001). In addition, participants were free to comment on their experience in a separate form.

4.2.2 Results

Since the main objective of this study was to explore the usability of immersive interactive storytelling, the most important results are constituted by the objective measures of user interaction during the immersive interactive storytelling experiments. However, one essential element of user acceptance is also the absence of adverse reactions: this is why we have studied cyber sickness for all subjects through the SSQ. Finally, it seemed appropriate to include at least one measure of presence.

The number of user actions per Interactive Narrative appears largely similar in Actor and Ghost mode (~60 actions), suggesting a comparable level of engagement. A one-way ANOVA test confirmed a similar engagement level for both modes, by demonstrating that the average number of actions did not differ significantly. This is also reflected in the average duration of the interactive narrative, between 5 and 6 minutes in total. It is worth commenting on the fact that interactivity can shorten the duration of the narrative, because participants were eager to interact with the story and thus users often tend to provide excessive emotional input to the main character Emma, prompting her to accelerate her decisions, leading to an anticipated ending (regardless of the exact ending itself). In practice, users have been able to extend the baseline story by 2 minutes or reduce it by nearly 3 minutes in certain cases.

Finally, with an average of 66% more distance covered in Ghost mode, it appears that users have spent more time exploring the environment. This can have several possible explanations: free exploration of the 3D stage as a way of enjoying the action, search for narrative objects supporting interaction, or simply lack of requirement to stay in close contact with Emma as in the Actor mode. This exploratory tendency is also visible in the high number of unproductive interactions, when users touch non-reactive object with their virtual hand.

Not surprisingly, the ratio of verbal to non-verbal interaction differs significantly in Actor mode, where users have to engage in dialogue with Emma as part of the Interactive Narrative, and in Ghost mode, where there is no such constraint. Non-verbal influences differ in nature between Actor mode and Ghost mode. For instance, in Ghost Mode, some participants did steal Emma's gift for Rodolphe (a book), and watched the consequences of their intervention, which from a system’s perspective, consisted in Emma’s “gift” action failing, and in the story world, resulted in Rodolphe expressing disappointment. It should be noted that users still used verbal interaction in Ghost mode, although its nature was very different from the novel’s dialogues. If we consider the overall duration of the Interactive Narrative, it appeared that while both IS paradigms seem to be equivalent in terms of user involvement (as evidenced by the number of actions), the Ghost mode corresponds to longer sessions, probably because of

the additional explorations by users.

After the experiment, we measured the user self-reported sense of spatial presence and engagement based on the ITC-SOPI Questionnaire (Lessiter, Freeman, Keogh, & Davidoff, 2001). The scores analysis confirms a high sense of spatial presence (70 %) as well as important level of user's engagement (74% on average). In their free reports gathered after the experiments, many users have spontaneously made reference to feelings of "being there": "A very good experience, I was surprised about how I felt in that environment. A sense of being there was strong", "I did lose track of time", "I felt immersed", "At a point I felt that I had involuntarily bumped into the table and moved back (even though I know it was not real)". The naturalness of the environment is slightly lower (64%) with average score of negative effect (50%), which are probably due to the relative simplicity of the environment's graphics and animations.

63.2% of the users contributed written comments after the experiments - of these 44.7 % included explicit positive statements such as "I enjoyed the experiment a lot", "I think the experiment was great", "It definitely was an interesting experiment", "Absolutely fantastic-could have done that all day", "I felt that I could really interact with Emma, which made the experience really interesting and pleasurable", "I was able to „steal" Emma's gift to Rodolphe, thus changing the outcome of the scene, I found it particularly enjoyable". Most of the 21% comments including negative aspects refer to disorientation; such as "I felt a little disoriented turning around while moving forward". Some comments also expressed certain preferences towards one particular interaction paradigm: "I preferred the role of Rodolphe as I felt there was a definite purpose. As a ghost I didn't feel really involved", "I enjoyed the first part of the story where I took the role of a character as that made me more comfortable in the environment as I had a role".

For more information on the study as well as on the results, please see (Lugrin, Cavazza, Pizzi, Vogt, & André, 2010).

4.3 Gaze-Based Interaction

Gaze-based interaction with EmoEmma was evaluated by comparing an interactive gaze model with a non-interactive gaze model. In the non-interactive mode, the agent's gaze model is parameterized in such a way that the agent seems randomly to look at the user or averts its gaze and the virtual character gazes on average for a period of 1 s (0-2 s) in any state. The interactive mode was realized as described in Section 3.1.2. Our objective was to find out whether the interactive mode had any positive effect on the user's social presence, engagement and interactional rapport compared to the non-interactive mode. Furthermore, we wanted to know to what extent the users' eye gaze behavior towards the agent followed patterns from human-human conversation.

4.3.1 Participants and Setting

For the experiment, we recruited 19 participants (2 females and 17 males) with a mean age of 25.3 (SD = 3.1). All participants were native speakers of German. First, the participants were placed in front of the projection screen. Then the eye tracker was calibrated. The participants were first informed about the background of the story. Then, they were told that they would enter the story in the role of Rodolphe who finds Emma alone in the salon and should try to engage her in a conversation. To exclude any side effects resulting from dynamically evolving stories of varying quality, we decided to use fixed story lines for the experiment. Thus, for the experiment, just EmoEmma's eye gaze behavior was automated. We do not consider fixed story lines as a major problem in this particular case since Emma's verbal utterances were carefully chosen so that the users could in general make sense of them. In addition, the scenario chosen - the user in the role of Rodolphe is expected to

approach Emma to start an affair with her - left the user with enough space for interpretation. In the experiment, Emma produced 12 turns pausing briefly (5-10 s) after each of them to give the user a chance to respond. Emma started with 'Hello Rodolphe, I am so delighted!' and the user could for example answer with 'Hello Emma, I feel just the same way!'. The whole process for each subject took about 20 minutes including the introduction to the story sequence whereby one interaction sequence took about 3 minutes. The order of the two gaze models (i.e. interactive and non-interactive) was randomized for each subject to avoid any bias due to ordering effects.

4.3.2 Social Presence, Engagement and Interactional Rapport

The objective of the study was to find out whether the different modes had any impact on the subjects' experience ratings. In particular, we used a post-questionnaire with a 9-point rating scale (from strongly disagree to strongly agree) to assess the subjects' sense of social presence (P), their level of rapport with the character (R), their engagement (E), the social attraction of the character (A) and the subjective perception of the story (S).

Measures. Social Presence (P). We assessed the subjects' sense of social presence using the items "I had the feeling that Emma was aware of me.", "I had the feeling of personal contact to Emma.", "Emma was impersonal." (reverse coded), and "Emma was reserved." (reverse coded).

Rapport with the Character (R). The level of rapport with the virtual character was measured using the items "I would have liked to continue the interaction with Emma.", "Emma's behavior was natural.", "I had the feeling that Emma reacted on me.", and "Emma's behavior was synchronous to mine.".

Engagement (E). We indexed the user's level of engagement with the following two items: "I enjoyed the first meeting with Emma." and "I found it easy to flirt with Emma.".

Social Attraction of the Character (A). The users' social attraction of the character was measured using "I had the feeling, that Emma was interested in me." and "Emma was sympathetic.".

Perception of the Story (S). The subjective perception of the story was assessed using the items "I would like to know how the episode with Emma continues.", "I had no problems to empathize with the part of Rodolphe.", and "I had the feeling to influence the story with my eye gaze.".

The significance analyses between the interactive gaze mode and the non-interactive mode were conducted using a paired two-tailed t-test. A look at Figure 11 reveals that all groups received more positive ratings for the interactive gaze model than for the non-interactive gaze model. The significance test reveals that the presence measure differs significantly between the interactive and non-interactive gaze mode (P: $t(75) = 2.6$; $p = 0.01$; $r = 0.29$). Also the rapport measure reveals significant differences between these two modes (R: $t(75) = 2.3$; $p = 0.02$; $r = 0.26$). However, the other measures did not reveal any significant differences (E: $t(37) = 1.6$; $p = 0.11$; A: $t(37) = 1.2$; $p = 0.25$; S: *$t(56) = 1.5$; $p = 0.15$).*

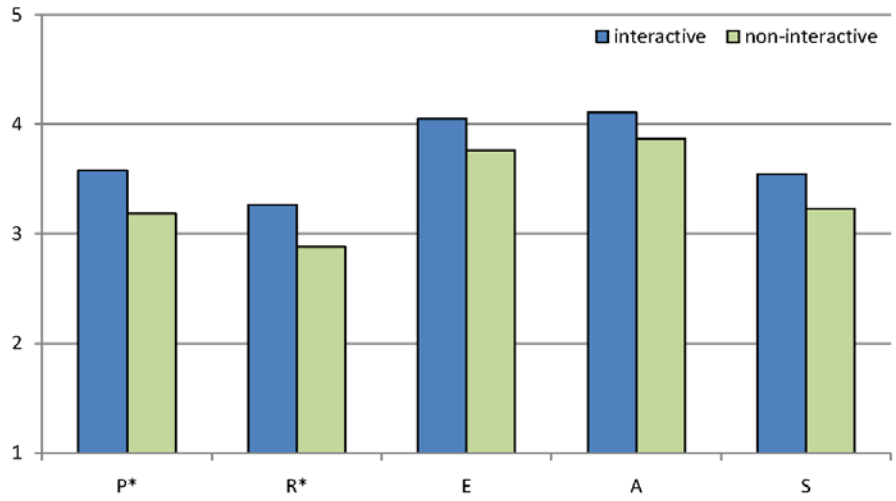

Figure 11: Results for the questions comparing the interactive and non-interactive gaze model while interacting with Emma (* $p < 0.05$)

4.3.3 Analysis of the Subjects Eye Gaze Behaviours

First of all, we investigated to what extent the subjects were looking at Emma while she was speaking or silent. This gives us evidence whether the user interacts with Emma in a similar way as they would do in human-human interaction. We calculated the fixation points from the raw eye gaze data using the algorithm presented in Section 2.4.1. Furthermore, we divided the scene into two areas. The first area covers the eyes of the virtual character and the second area the rest *of the scene.*

We found that independent from the gaze mode, the users were looking at Emma around 76% of the time in contrast to (Kendon, 1967) who found that in human-human interaction a human is looking on average 50% of the time at an interlocutor. Further, (Kendon, 1967) reports that this quote varies from 28% to 70% whereas we found a variation of 46% to 98%.

(Argyle & Cook, 1976) found that humans look about 75% at interlocutors while listening and 41% while speaking. Independent from the gaze mode, we found that users interacting with a virtual agent look about 81% of the time at the agent while listening and about 71% of the time at Emma while speaking. Although the users were in total much more looking at Emma, the relationship between listening and speaking remains comparable (i.e. the user looks at the interlocutor considerably longer when listening than when speaking) to human-human interaction. These findings are in line with a study conducted by (Rehm & André, 2005) and they ascribe them to the novelty effect of the agent.

Considering a multimodal gaze model that takes the user's eye gaze and speech into account, we analyze where the users are looking when they start and stop speaking. We expect findings that can be integrated in a multimodal interactive gaze model for a virtual character that enables the agent to detect whether a user plans to say or answer something or is expecting further advices from the system. In this way, an attentive system could recognize whether the current stimulus already suffices to expect an answer or feedback from the user or if the system needs to elaborate the current dialog part.

Figure 12 shows the gaze pattern when the users start speaking. We chose to analyze a 3.5 seconds interval, where we looked at the 3 seconds before the users started to speak and 0.5 seconds after the users started to speak. The users started to speak at t = 0 and we collected overall 430 utterances for this analysis. In Figure 12, three phases are shown: Emma speaks, pause and the user starts speaking. The pause after Emma speaks and the

user answers is on average 1.43 seconds (SD = 1.05). The vertical axis indicates the users' current gaze target, where 0 means the user looks away and 1 that the user looks at Emma's face. On average, the users looked significantly more at Emma while she was speaking than when the users started to answer, where they averted their gaze ($t(102) = 32.8$; $p = 0$; $r = 0.96$). The finding is not only statistically significant, but has also a large effect (r) and so indicates a substantive finding. (Morency, Christoudias, & Darrell, 2006) also found that users avert their gaze while thinking or answering.

In Figure 13 we plot the users' gaze pattern at the end of their utterance. We analyzed a 2.5 seconds interval, where we looked at 0.5 seconds before the users stop speaking and 2 seconds afterwards. The users stopped speaking at t = 0 and we collected overall 378 utterances from the users. The users started to looked significantly more often at Emma face after they stopped speaking ($t(123) = 6.2$; $p = 0$; $r = 0.49$). Looking at the gaze pattern in Figure 13 reveals that after the users end their utterance, their gaze behavior looks like an increasing saw tooth pattern. This means that they are rhythmically alternating their gaze between Emma's face and the rest of the virtual scene.

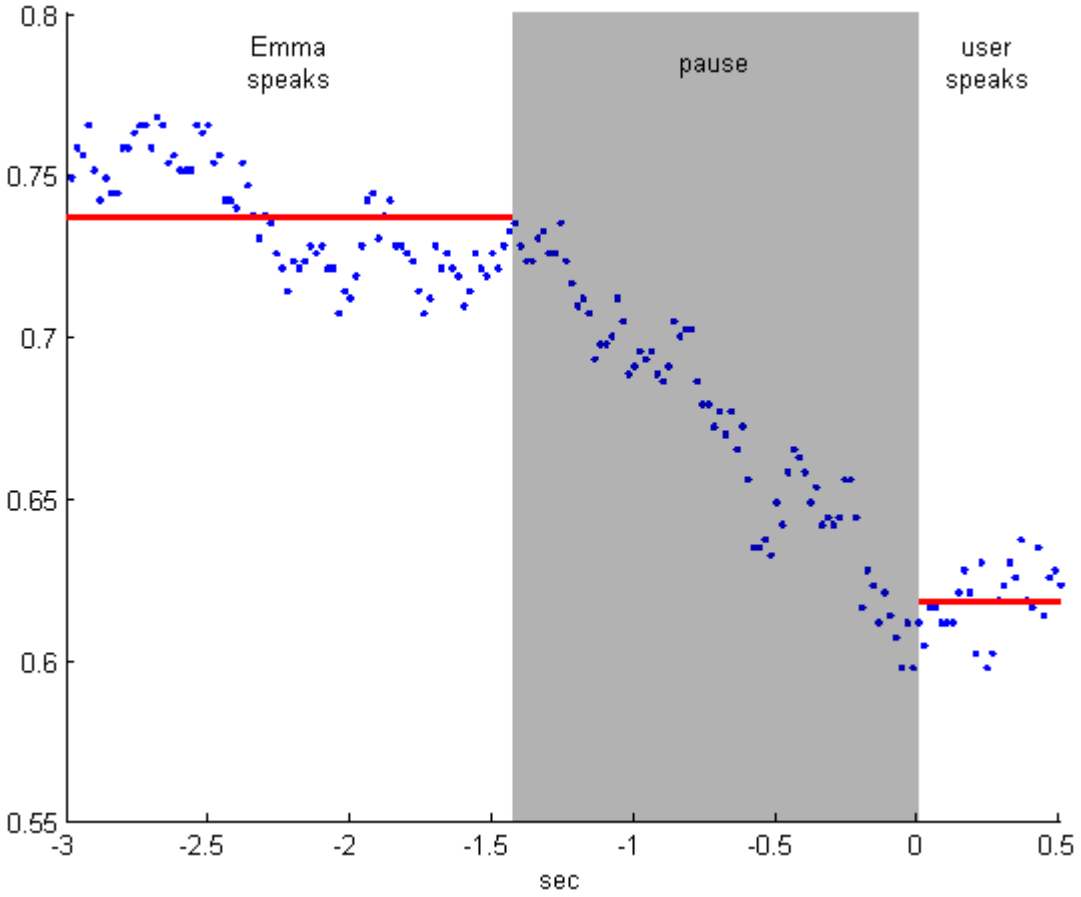

Figure 12: Eye gaze pattern before the users start speaking. The vertical axis indicates the gaze target (0 = looking away, 1 = looking at Emma, red line = average) during conversation. The user starts speaking at t = 0.

We found that the interactive gaze mode led to a better user experience compared to the non-interactive gaze mode. Indeed, the interactive gaze mode achieved a higher score for all items of a questionnaire measuring the user's sense of social presence, their level of rapport with the agent, their engagement, the social attraction of the character and the subjective perception of the story. These results correspond with our previous findings, where we

analyzed the interaction with a virtual character on eye gaze basis only. Additionally, we found that users adhere to patterns of gaze behaviors for speaker and addressee that are also characteristic of dyadic human-human interactions. However, they looked significantly more often to the virtual interlocutor than it is typical for human-human interactions.

A detailed description of the study and the results may be found in (Bee et al., 2010).

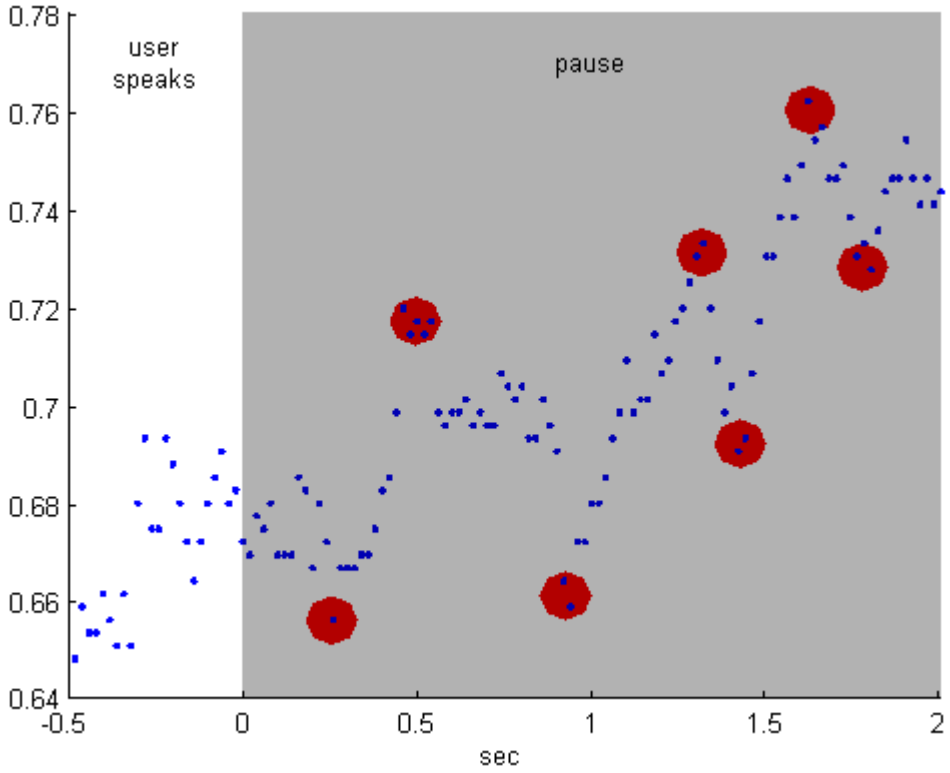

Figure 13: Eye gaze pattern after the users stopped speaking. The vertical axis indicates the gaze target (0 = looking away, 1 = looking at Emma). The *user stops speaking at t = 0.*

4.4 Natural Language Interaction in the Beergarden

Currently, we are preparing an empirical study to evaluate natural language interaction within the Virtual Beergarden. In particular, we are interested in the question of how the structure of the dialogue (character initiative vs. mixed initiative) influences the users' behaviour and their perception of the experience. To conduct the study, we will use a combination of questionnaires and behavioural measurements by means of log files and audio-visual recordings. The questionnaires will be based on the measurement toolbox developed by the University of Amsterdam (Vermeulen, Roth, Vorderer, & Klimmt, 2010) which includes scales covering the design dimensions discussed in Section 1.

- *Basic usability*

  Scales used to measure *system usability* and *correspondence of system capability with user experience*.

- *Narrative Engagement*

  Scales used to measure *curiosity, suspense, flow aesthetic appeal* and *user enjoyment*.

- *Narrative Immersion*

  Scales used to measure *presence* and *role adoption/identification.*

- *Narrative Flow*

  Scales used to measure *effectance.*

- *Conversational Flow*

  Scales used to measure *character believability.*

The behavioral measurements will be based on the Social Signal Interpretation Toolbox (Wagner, André, & Jung, 2009) developed by Augsburg University.

5. Conclusions

The technical framework presented in Section 2 has enabled us to explore a rich repertoire of interaction devices and modalities. Most of the presented components have been made publically available to the research community and attracted partially up to 100 downloads. Our focus was on the provision of communication means common in human-human conversation, such as speech, posture, gestures and gaze. First of all, these communication means matched the style of our IS scenarios best. Secondly, we aimed to keep the time required to get acquainted with an IS at a minimum. For this reason, we did not use, for example, a repertoire of symbolic gestures the user had to learn. To cope with the limitations of current speech technology, we designed scenarios in such a way that there was enough space of interpretation in case of unexpected system behaviour.

Various evaluations of the interaction devices and modalities revealed positive effects in terms of narrative engagement and narrative immersion. We observed that the ability of an IS system to understand and respond to emotional-socio cues already enables engaging and meaningful interactions. Furthermore, human participants showed behaviours when interacting with the characters of a story that follows patterns from human-human conversation. While the measurements used in the studies presented here, rely on questionnaires developed for games and 3D environments, our future studies will make use of the newly developed measurement toolbox developed within IRIS for IS systems.

References

Argyle, & Cook. (1976). *Gaze & Mutual Gaze.* Cambridge University Press.

Bee, N., Wagner, J., André, E., Vogt, T., Charles, F., Pizzi, D., et al. (2010). Discovering Eye Gaze Behavior during Human-Agent Conversation in an Interactive Storytelling Application. *12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010) .*

Cavazza, M., Pizzi, D., Charles, F., Vogt, T., & André, E. (2009). Emotional input for character-based interactive storytelling. In C. Sierra, C. Castelfranchi, K. S. Decker, & J. S. Sichman (Hrsg.), *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, May 10-15, 2009, Volume 1* (S. 313-320). IFAAMAS.

Colburn, A., Cohen, M., & Drucker, S. (2000). *The Role of Eye Gaze in Avatar Mediated Conversational Interfaces.* Microsoft Research.

Endraß, B., Boegler, M., Bee, N., & André, E. (2009). What Would You Do in Their Shoes? Experiencing Different Perspectives in an Interactive Drama for Multiple Users. In I. Iurgel, N. Zagalo, & P. Petta (Hrsg.), *Interactive Storytelling, Second Joint International Conference on Interactive Digital Storytelling, ICIDS 2009, Guimarães, Portugal, December 9-11, 2009. Proceedings. 5915*, S. 258-268. Springer.

Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., & Hagita, N. (2002). Messages embedded in gaze of interface agents –- impression management with agent's gaze. *CHI '02: Proc. of the SIGCHI conference on Human factors in computing systems* (S. 41-48). New York, NY, USA: ACM Press.

Garau, M., Slater, M., Bee, S., & Sasse, M. A. (2001). The impact of eye gaze on communication using humanoid avatars. *CHI '01: Proc. of the SIGCHI conference on Human factors in computing systems* (S. 309-316). ACM Press.

Gebhard, P., Kipp, M., Klesen, M., & Rist, T. (2003). *Authoring Scenes for Adaptive, Interactive Performances.* New York, NY, USA: ACM.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica , 26*, 22-63.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *3* (3), 203-220.

Kleinke, C. L. (1986). Gaze and Eye Contact: A Research Review. *Psychological Bulletin , 100* (1), 78-100.

Lee, S. P., Badler, J. B., & Badler, N. I. (2002). Eyes alive. *ACM Transactions on Graphics , 21* (3), 637-644.

Lessiter, J., Freeman, J., Keogh, E., & Davidoff, J. (2001). A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory. *Presence: Teleoper. Virtual Environ. , 10* (3), 282-297.

Lugrin, J.-L., Cavazza, M., Pizzi, D., Vogt, T., & André, E. (2010). Exploring the usability of immersive interactive storytelling. *Proceedings of the 17th ACM Symposium on Virtual Reality Software and Technology* (S. 103-110). New York, NY, USA: ACM.

Mehta, M., Dow, S., Mateas, M., & MacIntyre, B. (2007). Evaluating a conversation-centered interactive drama. *AAMAS '07: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems* (S. 1-8). New York, NY, USA: ACM.

Morency, L. P., Christoudias, M. C., & Darrell, T. (2006). Recognizing gaze aversion gestures in embodied conversational discourse. *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces* (S. 287-294). New York, NY, USA: ACM Press.

Nakano, Y. I., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. *ACL '03: Proc. of the 41st Annual Meeting on Association for Computational Linguistics* (S. 553-561). Association for Computational Linguistics.

Rehm, M., & André, E. (2005). From chatterbots to natural interaction - Face to face communication with Embodied Conversational Agents. *IEICE Transactions on Information and Systems, Special Issue on Life-Like Agents and Communication , 88-D* (11), 2445-2452.

Rowe, J., McQuiggan, S., & Lester, J. (2007). Narrative Presence in Intelligent Learning Environments. *Working Notes of the 2007 AAAI Fall Symposium on Intelligent Narrative Technologies* (S. 126-133). Washington, DC, USA: AAAI Press.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *ETRA '00: Proc. of the symposium on Eye tracking research & applications* (S. 71-78). ACM Press.

Schäfer, L. (2004). Models for Digital Storytelling and Interactive Narratives. *Cosign 2004.*

Shneiderman, B., & Plaisant, C. (2004). *Designing the User Interface: Strategies for Effective Human-Computer Interaction (4th Edition).* Pearson Addison Wesley.

Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2004). Where to look: a study of human-robot engagement. *IUI '04: Proc. of the 9th international conference on Intelligent user interfaces* (S. 78-84). ACM Press.

Velten, E. (1968). A laboratory task for induction of mood states. *Behavior Research & Therapy , 6* (4), 473-482.

Vermeulen, I. E., Roth, C., Vorderer, P., & Klimmt, C. (2010). Measuring User Responses to Interactive Stories: Towards a Standardized Assessment Tool. In R. Aylett, M. Y. Lim, S. Louchart, P. Petta, & M. Riedl (Hrsg.), *Interactive Storytelling - Third Joint Conference on Interactive Digital Storytelling, ICIDS 2010, Edinburgh, UK, November 1-3, 2010. Proceedings. 6432*, S. 38-43. Springer.

Vinayagamoorthy, V., Garau, M., Steed, A., & Slater, M. (2004). An Eye Gaze Model for Dyadic Interaction in an Immersive Virtual Environment: Practice and Experience. *Computer Graphics Forum , 23* (1), 1-11.

Vogt, T., André, E., & Bee, N. (2008). EmoVoice - A framework for online recognition of emotions from voice. *Proc. of Workshop on Perception and Interactive Technologies for Speech-Based Systems.* Springer.

Wagner, J., André, E., & Jung, F. (2009). Smart sensor integration: A framework for multimodal emotion recognition in real-time. *Affective Computing and Intelligent Interaction (ACII 2009).*