Laugh-aware Virtual Agent and its Impact on User Amusement

Radosław Niewiadomski
TELECOM ParisTech
Rue Dareau, 37-39
75014 Paris, France
niewiado@telecom-paristech.fr

Jennifer Hofmann
Universität Zürich
Binzmuhlestrasse, 14/7
8050 Zurich, Switzerland
j.hofmann@psychologie.uzh.ch

Jérôme Urbain
Université de Mons
Place du Parc 20
7000 Mons, Belgium
jerome.urbain@umons.ac.be

Tracey Platt
Universität Zürich
Binzmuhlestrasse, 14/7
8050 Zurich, Switzerland
tracey.platt@psychologie.uzh.ch

Johannes Wagner
Universität Augsburg
Universitätsstr. 6a
86159 Augsburg, Germany
wagner@hcm-lab.de

Bilal Piot
Supelec
Rue Edouard Belin, 2
57340 Metz, France
bilal.piot@supelec.fr

ABSTRACT

In this paper we present a complete interactive system enabled to detect human laughs and respond appropriately, by integrating the information of the human behavior and the context. Furthermore, the impact of our autonomous laughter-aware agent on the humor experience of the user and interaction between user and agent is evaluated by subjective and objective means. Preliminary results show that the laughter-aware agent increases the humor experience (i.e., felt amusement of the user and the funniness rating of the film clip), and creates the notion of a shared social experience, indicating that the agent is useful to elicit positive humor-related affect and emotional contagion.

Categories and Subject Descriptors

H.5.2 [**Information Technology and Systems**]: Information Interfaces and Representation (HCI)User Interfaces[Graphical user interfaces]; H.1.2 [**Information Technology and Systems**]: Models and PrinciplesUser/Machine Systems[Human factors]

General Terms

Algorithms, Experimentation

Keywords

laughter detection, laughter synthesis, virtual agents

1. INTRODUCTION

Laughter is a significant feature of human communication, and machines acting in roles as companions or tutors should not be blind to it. So far, limited progress has been made towards allowing computer-based applications to deal with laughter. In consequence, only a few multimodal systems exist that use laughter in the interaction. In this paper we present a complete laughter-aware interactive system that is able to detect human laughs and to respond appropriately (*i.e.*, laugh with right timing and intensity) to the human behavior and the context. For this purpose we adapted and integrated several existing open-source analysis and synthesis tools. On top of them we developed user laughter detection and laughter intensity estimation modules as well as audiovisual laughter synthesis. We also provided a decision making module that controls the agent behavior. All these components work in real-time. In addition, we built an interactive scenario for our autonomous laughing agent. In this scenario, the user watches a pre-selected funny stimulus (*i.e.*, film clip, cartoon) alongside the agent. The agent is able to laugh, reacting to both, the context (i.e., punch lines in the film clip) and the user's behavior. The impact of the interaction with the laughing agent and its expressive behavior is assessed by evaluation questionnaires covering emotional, motivational, and cognitive aspects of the experience, as well as beliefs and opinions towards the agent.

2. RELATED WORK

Urbain et al. [1] have proposed the AVLaughterCycle machine, a system able to detect and respond to human laughs in real time. With the aim of creating an engaging interaction loop between a human and the agent they built a system capable of recording the user's laugh and responding to it with a similar laugh. The virtual agent response is automatically chosen from an audiovisual laughter database by analyzing acoustic similarities with the input laughter. This database is composed of audio samples accompanied by the motion capture data of facial expressions. While the audio content is directly replayed, the corresponding motion capture data are retargeted to the virtual model. Shahid et al. [2] proposed Adaptive Affective Mirror, a tool that is able to detect user's laughs and to present audiovisual affective feedback, which may elicit more positive emotions in the user. In more details, Adaptive Affective Mirror produces a distortion of the audiovisual input using real-time

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.*

graphical filters such as bump distortion. These distortions are driven by the amount and type of user's laughter that has been detected. Fukushima et al. [3] built a system able to increase users' laughter reactions. It is composed of a set of toy robots that shake heads and play preregistered laughter sounds when the system detects the initial user laughter. The evaluation study showed that the system enhances the users' laughing activity (*i.e.*, generates the effect of contagion). Finally, Becker-Asano et al. [4] studied the impact of auditory and behavioral signals of laughter in different social robots. They discovered that the social effect of laughter depends on the situational context including the type of task executed by the robot, verbal and nonverbal behaviors (other than laughing) that accompany the laughing act [5].

3. SCENARIO

For the purpose of building a laugh-aware agent we searched for an interaction scenario where laughter-based interaction appears to be natural and realistic in both human-human (for training the system) and human-machine conditions. Thus, we opted for a scenario that implies telepresence. Our scenario involves two subjects watching a funny stimulus (i.e., film clip). Importantly, they do not share the same physical space: they watch the same content simultaneously on two separate LCD displays. They can see the partner's reactions in a small window because a view of the other person is placed on the top of the displayed content. This scenario corresponds to very common situation in real life when someone wants to share interesting content over the web. For this reason teleconference systems such as Ekiga or Skype are often used.

We adapted this scenario to human-virtual agent interaction. It has many advantages: the situational context is limited to the presented context, interaction should be mainly based on laugh episodes and is possible without using speech recognition, which is often a bottleneck for current interactive systems. The other advantage of this scenario is that it allows us to easily alter the interaction conditions. This is important as we want to evaluate the impact of the laugh-aware agent for interaction. For this purpose we introduce three variants of our scenario:

- "fixed speech" (FSC): the agent is verbally expressing amusement at pre-defined time slots,
- "fixed laughter" (FLC): the agent is expressing amusement through laughs at pre-defined times slots,
- "interactive laughter" (ILC): the agent is expressing amusement through laughter, in reaction to both the content and the human's behavior.

4. SYSTEM ARCHITECTURE

To realize our scenario we have built a laugh-aware autonomous agent able to analyze the human laugh behavior and answer appropriately in real-time. It is composed of several modalities (see Figure 1). We can distinguish 3 types of components: input components, decision components and output components.

The input components are responsible for multimodal data acquisition and real-time laughter-related analysis. They include laughter detection from audio features and input laughter intensity estimation. Data is collected by a Microsoft Kinect, video (RGB, 30fps, 640x480) and audio (16 kHz, 16 bit, mono). Two decision components are used to control the agent audiovisual response. The first one (Dialog Manager) receives the information from the input components (*i.e.*, laughter likelihoods and intensity) as well as contextual information and it generates a high-level information on laughter response (i.e., its duration and intensity). The second component, Laughter Planner, controls the details of the expressive pattern of the laughter response. In the current version it chooses the appropriate audiovisual episode from the lexicon of pre-synthesized laughter samples and it encodes it into Behavior Markup Language (BML)[1]. This two-step decision process allows us to separate the model of interaction and the agent specific characteristics. While the Dialog Manager is responsible for high level decisions that are independent of the agent or its embodiment, the Laughter Planner plans the audiovisual response taking into consideration the agent characteristics such as gender or available modalities etc. Finally, two output components are responsible for the audiovisual laughter synthesis and visualization.

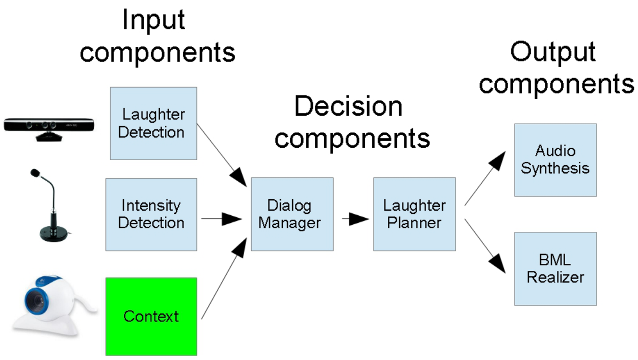

Figure 1: Overall architecture of laugh-aware agent

All the components work in real-time. To ensure fast and efficient communication between the different modules, we use the message-oriented middleware called ActiveMQ^{TM} which supports multiple operating systems and programming languages. Each component can read and write to some specific ActiveMQ topics. For this purpose we defined a hierarchy of message topics and for each topic the appropriate message format. Simple data (such as input data) were coded in simple text messages in string/value tuples, so called *MapMessage*s, while the description of the behavior to be displayed by the agent was coded in standard XML-like language.

4.1 Input Components

To facilitate the multimodal data processing and the synchronization between the different signals, we use existing software called Social Signal Interpretation (SSI) [6]. It is used to collect, synchronize and process multimodal data from different data sources such as Kinect, microphone or physiological sensors. For the purpose of our laugh-aware agent we developed new modules in SSI: audio laughter detection and laughter intensity estimation. We also use SSI to include information about the context of the interaction.

[1] http://www.mindmakers.org/projects/bml-1-0/wiki/Wiki?version=10

4.1.1 *Laughter detection*

To find which feature set is appropriate for laughter detection a large scale experiment was conducted. The following speech-related low-level features were selected as most promising candidates: Intensity, MFCCs, Pitch, PLPs. On these the following 11 groups of functionals were tested: Zero-Crossings, DCT (Direct Cosine Transform) Coefficients, Segments, Times, Extremes, Means, Onsets, Peaks, Percentiles, Linear and Quadratic Regression, and Moments. These features were extracted by the openSMILE feature extraction toolkit [7], integrated into SSI. The experiment was run on 19 sessions of the SEMAINE corpus [8], which were manually annotated into laughter and speech samples and distributed in a training and test set, while it was ensured that samples of the same user would not occur in both sets. For classification we applied Support Vector Machines.

Results suggest that most reliable results are achieved using Intensity and MFCCs, while adding Pitch and PLP features did not improve results on the studied corpus. Among the functionals, Regression, Moments, Peaks, Crossings, Means and Segments are considered to carry most distinctive information. In the best case an overall accuracy of 88.2% at an unweighted average recall of 91.2% was obtained.

The developed laughter detection framework was then tuned to our interactive scenario and input component. 20 subjects were recorded with our scenario settings while watching the stimulus video. Laughter annotations were used to re-train the laughter detector described above. The obtained laughter model was finally combined with a silent detection to filter out silent frames in the first place and classify all remaining frames into laughter or noise. The frame size was set to 1 second with an overlap of 0.8 second, *i.e.* a new classification is received every 200ms.

4.1.2 *Laughter intensity*

Knowing the intensity of the incoming laugh is important information to determine the appropriate reaction. We built and integrated into SSI a laughter intensity estimation module that takes into consideration the audio characteristics of user laughs. For this purpose we used some episodes of the freely available AVLC database [9] that contains nearly 1000 audiovisual spontaneous laughter episodes. 49 acoustic laughs, produced by 3 subjects, have been continuously annotated in intensity by one labeler. Looking at the intensity curves and the dynamics of standard audio features (energy, pitch, MFCCs, ...), a simple function to automatically estimate the intensity has been designed: the intensity curve is obtained by a linear combination between the maximum pitch and the maximum loudness values over a sliding 200ms window, followed by median filtering to smooth the curve. A comparison between the manual intensity annotation and our automatic estimation is presented on Figure 2. It can be seen that the overall trend is followed, even though there are differences, mostly at the edge of the manually spotted bursts, and the manual curve is smoother than the automatic one.

4.1.3 *Context information*

The context in our scenario is defined according to the ranging of funniness of the displayed content. For this purpose we asked 14 labelers to give a continuous evaluation of the funniness of the presented content. The results of the analysis (for details see Section 5.2.1) of their responses are then synchronized with the displayed content. The current information about funniness is available through ActiveMQ to all the components. It is used by the Dialog Manager (see Section 4.2.1) in ILC condition and directly by the Laughter Planner (see Section 4.2.2) in FLC and FSC conditions.

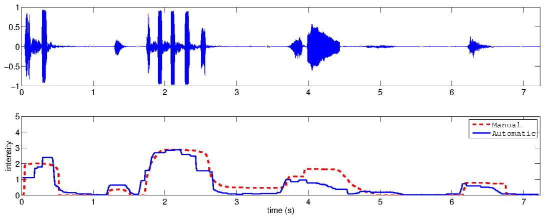

Figure 2: Example of laughter continuous intensity curve. Top: waveform; Bottom: manual and automatic intensity curves.

4.2 Decision components

The laughter-enabled decision making module aims at deciding, given the information from the input components, when and how to laugh so as to generate a natural interaction with human users. It is composed of two layers: the Dialog Manager which is responsible for high-level decisions about when to generate laughter and the Laughter Planner that plans the details of laughter audiovisual response.

4.2.1 *Dialog Manager*

The input I received by the Dialog Manager at each 200ms time frame is a vector ($I \in [0,1]^k$: each feature being normalized) where k is the number of features (5 in our case). The output O produced at each time frame is a vector ($O \in [0,1] \times [0, \mathrm{time}_{\mathrm{max}}]$) where the first dimension codes the laughter intensity and the second dimension codes the duration of the laugh.

To train the Dialog Manager, we used data recorded with the participation of two user dyads. The users (named $P1$ and $P2$) watch simultaneously, but in separate rooms, the same stimulus video: $P2$ is viewable by $P1$ and is, in our model, considered as playing the role of the virtual agent. The objective is to find a decision rule such that the virtual agent will imitate $P2$. To do so, an hybrid supervised/unsupervised learning method has been used. In a first stage, the inputs are clustered in N classes with a k-means method and the outputs are clusted in M classes via a GMM method. Secondly, a matching π between input and output clusters is computed thanks to a k-nearest neighbors algorithm (k-nn). K-nn is a classification method associating inputs and outputs using a majority vote among the k nearest examples in the training set.

Because we want to introduce some variability into the generated laughs, the actual output of the dialogue manager is randomly drawn from the distribution of the examples in the output cluster (instead of the centroid). This is why we use a GMM method for clustering the outputs: in each cluster l, the distribution of samples can be seen, in the 2-dimensional intensity-duration plane, as a Gaussian of law $\mathcal{N}(\mu_l^O, \Sigma_l^O)$. Therefore, to obtain an output, it is sufficient to sample an element O of law $\mathcal{N}(\mu_l^O, \Sigma_l^O)$. This operation is called the output generation.

Figure 3 summarizes the Dialog Manager operations: at each time frame the input vector I is associated to the closest input cluster $I^C \in \{1, \ldots, N\}$, then the decision rule π gives the output cluster $\pi(I^C) \in \{1, \ldots, M\}$, finally the output O is chosen in the output cluster $\pi(I^C) \in \{1, \ldots, M\}$ via the output generation.

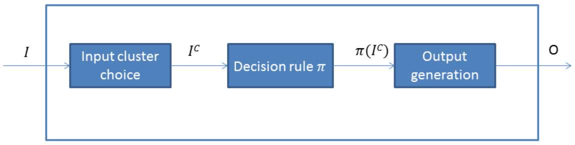

Figure 3: Dialog Manager functioning

4.2.2 *Laughter Planner*

In our laugh-aware agent the Dialog Manager is followed by the Laughter Planner, which is adapting the outputs of the Dialog Manager to the constraints of the synthesis modules. The Laughter Planner module can work in three different conditions (FSC/FLC/ILC). In the first two conditions (FSC and FLC), the Laughter Planner receives the information about the context (time of funny event, see Section 5.2.1) and it sends to the agent the pre-scripted BML verbal (FSC) or nonverbal (FLC) reaction to be displayed. The list of these behaviors was chosen manually.

In ILC condition the behavior of the agent is flexible as it is adapted to the user and the context. The Laughter Planner receives from the Dialog Manager the information on duration and intensity of laughter responses and using these values it chooses one laugh episode from the library that matches the best both values, i.e. the less distant (in the sense of a weighted Euclidean metric) episode i, (int_i, dur_i) that minimizes the value of: $\sqrt{(3 * \Delta(int))^2 * (\Delta(dur))^2}$. At the moment, the synthesis components do not allow for interruptions of the animation. Once it is chosen, the laugh episode has to be played until the end. During this period the Laughter Planner does not take into account any new information coming from the Dialog Manager. All the episodes start and end with a neutral expression.

4.3 Output components

The role of output components is to synthesize and display the agent laugh-based behaviors. At the moment laughter synthesis is realized off-line. The laughter episodes are synthesized using a freely available database of human laughter (AVLC corpus [9]). We synthesize separately the acoustic and the visual modalities, using the original audiovisual signals (with synchronized audio and video flows). All the re-synthesized episodes are stored in the agent lexicon, and can then be displayed in real-time. In the following sections we present details of laughter audio and visual synthesis.

4.3.1 *Acoustic laughter synthesis*

Given 1) the lack of naturalness resulting from previous attempts at the acoustic synthesis of laughter [10], 2) the need for high level control of the laugh synthesizer and 3) the good performance achieved with Hidden Markov Model (HMM) based speech synthesis [11], we decided to investigate the potential of this technique for acoustic laughter synthesis. We opted for the HMM-based Speech Synthesis System (HTS) [12], as it is free and widely used in speech synthesis and research. Explaining the details of speech synthesis with HMMs or HTS goes beyond the scope of this paper. In short, the standard HTS process has been followed. HMMs have been trained with the help of the AVLC database and its phonetic annotations [13]. Some modifications have been made to the data annotations to better exploit the potential of HTS: a syllable annotation layer has been introduced, and phones have been grouped in 8 broad phonetic classes—namely: fricatives, plosives, vowels, humlike (including nasal consonants), glottal stops, nareal fricatives (noisy respiration airflow going through the nasal cavities), cackles (very short vowel similar to hiccup sound) and silence—to increase the number of samples available to train each phonetic model. In addition, some standard parameters of HTS have been tuned to our laughter voices (e.g., the fundamental frequency boundaries).

After training laughter synthesis HMMs, we can produce acoustic laughs from acoustic laughter transcriptions. It is worth noting that there is currently no module to generate such laughter phonetic transcriptions from high-level instructions (*e.g.*, a type of laughter, its duration and its intensity). For the moment, only existing acoustic transcriptions from the AVLC database are synthesized, in synchrony with their visual counterparts (see following section).

4.3.2 *Visual laughter Synthesis*

For the visual synthesis we use an existing BML Realizer that can be driven by both anatomically inspired facial behavior description based on the Facial Action Coding System (FACS) [14] or low-level facial animation parameterization (FAPs) that is a part of MPEG-4 standard [15] for facial animation. To create the lexicon of laughter facial expressions we use both, procedural animation based on the manual annotation of action units (AUs) on the video, as well as synthesis based on automatic facial action detection from landmarks.

In the first approach, using FACS and viewing digital-recorded facial behavior at frame rate and in slow motion, a selection of twenty pre-recorded laughter events were coded by certified FACS coders. These codes were automatically converted in BML, in order to be displayed by BML Realizer.

In the second approach we estimate the facial animation parameters (FAPs) for each frame of animation by using an open-source face tracking tool – FaceTracker [16]. It uses a Constrained Local Model (CLM) fitting approach track facial landmark localizations. It can detect 66 facial landmark coordinates within real-time latency. FAPs are estimated as a distance between facial landmarks and neutral face landmarks where a default neutral face model is created with the help of 50 neutral faces of different persons. Some landmarks can be directly mapped to corresponding FAPs because they have exactly the same positions on the face while other FAPS are computed using linear interpolations between two or more landmarks. The landmark coordinates produced by FaceTracker are observed as noisy due to the discontinuities and outliers in each facial point localization. In order to smooth the face tracking parameters, a temporal regression strategy is applied on individual landmarks by fitting 3rd order polynomial coefficients on a sliding window, where the sliding window size is 0.67 seconds (*i.e.*, 16 frames) and sliding rate is 0.17 seconds (*i.e.*, 4 frames).

5. EVALUATION

Humor research has shown that the presence of another person influences cognitive and emotional responses to humor stimuli. Chapman [17] showed that a companion's laughter increased the displayed laughter, smiling, and ratings of funniness of humorous material. Consequent studies on the role of others for the laughing behavior included factors like the presence of a laughing versus a non-laughing model, proximity, crowding, eye contact, seating position, age difference between participants, and whether groups of strangers or friends were tested, as well as the role of state cheerfulness [17, 18]. Several of these factors have been shown to enhance the frequency and/or duration of smiling and laughter. Also, the potential facilitation of amusement through virtual agents and their expression of humor appreciation in face and voice has been claimed (see e.g., [19]). In consequence, the laughing virtual agent, similarly to the human companion, might be able to facilitate emotional contagion of amusement in an interactive setting. Therefore the purpose of this evaluation was to investigate the laughing agent and its impact on the humor experience. Our evaluation was organized according to three conditions (FSC/FLC/ILC) of the scenario proposed in Section 3 and entails a user watching a funny film alongside the laughing agent. The set of hypothesis concerning the perception of agent and interaction is treated with subjective measures assessed by self-report questionnaires.

In more details, this first evaluation focuses on two aspects. Firstly, the quality of the agent's laughter (i.e., the naturalness of the agent's laughter in face and voice, as well as the timing and adequacy of response) were evaluated. Here, it is assumed that the perceived naturalness is comparable over all conditions of the experimental scenario, as the methods used for generating the agent's responses are the same. Secondly, the impact of the agent behavior on the humor experience of the user when watching the funny film clip is investigated. More specifically, in the two fixed conditions (FSC, FLC), the agent expresses amusement but does so independent of the user. In the interactive condition, the agent reacts to the user's laughter with laughter, which can be interpreted as the agent responding in the most natural way to the user's behavior, creating a rapport, and facilitating emotional contagion of amusement when watching a funny video. Compared to the fixed conditions, where the agent acts independently, the interactive condition allows for a mutual response pattern in user and agent. Therefore, it was assumed that the interactive condition would lead to higher scores in felt amusement (H1) and the agent's laughter should be perceived more contagious (H2) compared to the two fixed conditions. Also, more intense social connection should be felt in the interactive condition compared to the fixed ones (H3).

5.1 Evaluation Questionnaire

To evaluate the quality of the interaction with the virtual agent, the naturalness of the virtual agent and felt emotions, an evaluation questionnaire was utilized. It consists of four broad dimensions targeting the evaluation of the naturalness, emotions, cognitions towards the agent, social aspects, as well as general questions. Items are formulated by using the term "avatar", as it is more easily understood by naïve participants. For the purpose of this paper, the relevant item clusters for the stated hypothesis on a laughing virtual agent were selected. Five items target the judgment of quality of the virtual agent's expressive behaviour (5 items; *e.g.*, "the laughter of the virtual agent was very natural", "The avatar's facial expression matched the vocal expressions"), forming a scale on naturalness (Cronbach's Alpha = .83). Experienced amusement, as the specific humor related affect was chosen from the broader scale on positive experience (4 items; e.g., "the virtual agent increased my amusement"), building a subscale "amusement" (Cronbach's Alpha = .79). The contagiousness of the agent's laughter ("The laughter of the avatar was contagious"), is represented by one item. The social presence/connection (6 items, "I felt company when interacting with the avatar", "I felt connected to the avatar") was assessed with the respective scale of the questionnaire (Cronbach's Alpha = .81). All items are judged on a seven point Likert-scale (1 = strongly disagree to 7 = strongly agree). In the three open questions, participants can express any other thoughts, feelings or opinions they would like to mention, as well as describing what they liked best/least.

5.2 Conditions

The three conditions (fixed speech (FSC), fixed laughter (FLC), interactive laughter (ILC)) differ in the degree of expressed appreciation of the clip (amusement) in verbal and non-verbal behavior, as well as degree of interaction with the user's behavior. In the fixed speech and fixed laughter conditions, the agent would be acting independent of the user, but still be signaling appreciation towards the funny film. In the interactive condition, the agent would be responding to the user's behavior. In other words, only the contextual information is used in the FSC and FLC conditions, while the input and decision components (see Sections 4.1 and 4.2.1) were active in the interactive condition.

5.2.1 Time points selection in FSC and FLC

The pre-defined times were chosen from the stimulus video. Firstly, 14 subjects (3 females, 11 males) watched the video material and annotated the funniness to it on a continuous funniness rating scale (ranging from "not funny at all" to "slightly funny", to "funny", to "really funny" to "irresistibly funny"). Averaged and normalized funniness scores were computed over all subjects, leading to sections with steep increases in funniness (apexes; see Figure 4) over the video. Secondly, trained raters assigned "punch lines" to the stimulus material, basing on assumptions of incongruity-resolution humor theory. Whenever the incongruous situation/prank was resolved for the subject involved, and amusement in the observer would occur from observing the resolution moment, a peak punch line was assigned. Punch lines were assigned for the first punch line occurring and the last punch line occurring in a given clip. When matching the continuous ratings with the punch lines, it was shown that the funniness apexes did cluster within the first and last punch lines for all subjects and all pranks, apart from one outlier. From these analyses, 8 funny moments have been selected: 2 for each of the 3 long pranks, 1 for the other 2. Pre-defined time points were controlled for a 1.5s delay in the rating/recording, due to reaction latency of the subjects and motor response delay.

5.2.2 Fixed Speech

In the FSC condition, the agent expressed verbal appre-

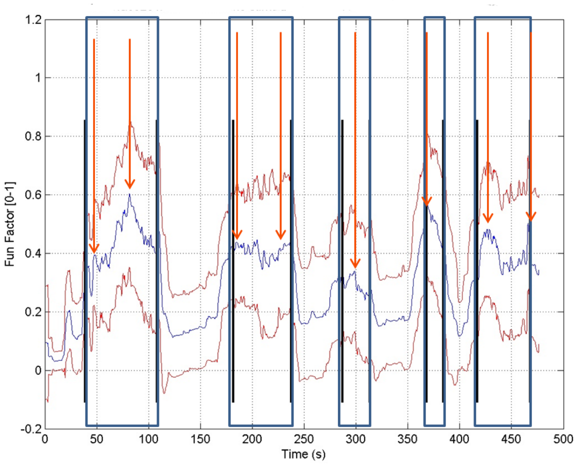

Figure 4: Continuous funniness ratings (means in blue and standard deviations in red) over the stimulus video for 14 subjects and expert assigned punch lines (first and last, in blue) to each clip. Red arrows indicate time points for fixed responses.

ciation in 8 short phrases (*e.g.*, “oh, that is funny”, “I liked that one”, “this is great”, “how amusing”, “I wonder what is next”) at pre-defined times. The verbal responses were rated for intensity on a four-point scale and matched to the intensity scores of the pre-defined time points.

5.2.3 Fixed Laughter

In the FLC condition, the agent laughed at pre-defined times during the video. The times were the same as the time points in the fixed speech condition. The agent displayed 8 laughs which varied in intensity and duration, according to the intensity ratings of the pre-defined time points.

A laughter bout may be segmented into an onset (*i.e.*, the pre-vocal facial part), an apex (*i.e.*, the period where vocalization or forced exhalation occurs), and an offset (*i.e.*, a post-vocalization part; often a long-lasting smile fading out smoothly; see [20]). Therefore, the onset was emulated by an independent smiling action just before the laughter (apex) would occur at the fixed time. The offset of the laughter was already integrated in the 8 chosen laughs.

5.2.4 Interactive Condition

In the ILC condition, the agent was using two sources of information to respond to the user: the continuous funniness ratings to the clip (context, shown in Figure 4) and the user’s acoustic laughter vocalizations. The Dialog Manager was receiving these two information flows and continuously taking decisions about whether and how the virtual agent had to laugh, providing intensity and duration values of the laugh to display. These instructions were then transmitted to the audiovisual synthesis modules.

5.3 Procedure of the evaluation study

The study consisted of the filling in of questionnaires (approximately 30-45 minutes) and a session of 30 minutes. No further information on the aims of the study was given. At the experimental session, users were assigned to one of the three conditions. Then, they were asked to fill in some initial questionnaires which were part of a broader assessment. Afterwards they were asked to sit in front of a television screen on a cushion about 1m from the screen. A camera allowed for the frontal filming of the head and shoulder and upper body of the user. They were also given headphones to hear the virtual agent. The experimenter explained that the user was asked to watch a film together with virtual agent and that the experimenters would leave the room when the experiment started. Once the experimenters left the room, the agent greeted the user and subsequently, the video started. After the film, the post measures/evaluation questions were filled in. After all questionnaires were completed, the user was debriefed and asked for written permission to use the obtained data. All users agreed to their data being used.

The following setup was used in this experimental session. Two LCD displays were used: the bigger one (46”) was used to display the stimuli (the funny film). The smaller (19”) LCD display placed on the right side of the big one was used to display the agent (a close-up view of the agent with only the face visible was used). The stimulus film consisted of five candid camera pranks with a total length of 8 minutes. The clips were chosen by one expert rater who screened a large amount of video clips (approximately 4 hours) and chose five representative, culturally unbiased pranks sections of approximately 1 to 2 minutes length. All pranks were soundless and consisted of incongruity-resolution humor.

6. RESULTS

Twenty-one users (13 males; ages ranging from 25 to 56 years, $M = 33.16$, $SD = 8.11$) volunteered to participate. Four users were assigned to the fixed speech condition, 5 to the fixed laughter condition and 11 to the interactive condition. Next, the results for the analysis of the naturalness and the stated hypothesis H1 to H3 are presented (see Table 1). Due to the low sample size, the two fixed conditions were compiled ($N = 9$) and compared to the interactive condition ($N = 11$), as the degree of interaction was the main focus of the experimental variation.

6.1 Naturalness of the agent laughter

In respect to the naturalness of the agent’s laughter, five items were rated by the users (e.g., “the avatars facial expressions matched the laughing sounds”). The item means ranged from $M = 2.22$ ($SD = 1.20$; fixed condition; “The behavior of the avatar was very natural”) to $M = 4.27$ ($SD = 1.79$; interactive condition; “The way the avatar laughed was similar to human laughter”). The means of the items indicate that the users choose values around the mid-point of the rating scale and also beyond, indicating that the laughter of the agent was not rated “extremely natural” in any condition. Next, the five items were aggregated to form a scale on naturalness and a oneway ANOVA (condition as the factor and naturalness as the dependent variable) was performed to reveal any mean differences between the fixed and interactive conditions. Results show that the two group means did not differ in the rated naturalness, F (1, 19) = 2.90, $p = .137$.

6.2 Amusement, Emotional Contagion and Social Experience

Concerning the hypothesis H1, four items of the questionnaire forming the subscale amusement were aggregated. A oneway ANOVA was computed with condition as independent (fixed vs. interactive) and the aggregated amusement score as dependent variable. The interactive condition led to higher felt amusement in the users, compared to the fixed conditions, F (1, 20) = 3.10, $p < .05$, one-tailed, $\eta^2 = 0.146$, showing that the virtual agent contributed to the humor experience.

In respect to the hypothesis H2 of emotional contagion of amusement (laughter contagion), a oneway ANOVA with condition as independent variable (fixed vs. interactive) and the item on laughter contagion was computed. In line with the expectations, the interactive condition yielded higher scores on the item "the laughter of the avatar was contagious" than the fixed conditions, F (1, 19) = 6.35, $p < .05$, $\eta^2 = 0.261$.

In respect to the hypothesis H3, targeting the experienced social presence with the agent, the six items of the questionnaire were aggregated and used as the dependent variable in a oneway ANOVA (condition as independent variable). Although numerically rating social presence and experience higher in the interactive than the fixed conditions, results show that the difference between the groups failed to reach statistical significance (F [1, 20] = 2.68, $p = .119$).

Table 1: Evaluation results (standard deviation in brackets, significant differences in bold).

	Fixed	*Interactive*	*F (1, 19)*
Naturalness	2.50 (1.10)	3.48 (1.40)	2.90
H1: Amusement	2.28 (0.96)	**3.30 (1.51)**	**3.10**
H2: Contagion	1.67 (0.50)	**3.27 (1.85)**	**6.35**
H3: Presence	2.07 (0.69)	2.89 (1.36)	2.68

6.3 Open Answers

Out of the 21 users, 14 gave answers to the question of what they liked least about the session. Half of the users mentioned that the video was not very funny or would have been funnier with sound. Two users mentioned that they could not concentrate on both, the virtual agent and the film. 17 users responded to what was liked best about the session. Best liked was the laughter of the virtual agent through the headphones (it was considered amusing and contagious; three nominations), the video (five nominations), the set up (four nominations) and one person stated: "It was interesting to see in what situations and in what manner the virtual agent responded to my laughter and to funny situations respectively" (subject 12).

7. DISCUSSION

Overall, the perceived naturalness of the agent's laughter was comparable over the conditions, in line with the expectations. In respect to the hypothesis H1 to H3, the results indicate that H1 (amusement) and H2 (laughter contagion) were confirmed, while H3 (social experience) failed to reach statistical significance and was therefore rejected.

Concerning the naturalness of the displayed laughter, results indicate that users did not rate the laughter produced by our agent as very natural. This finding was independent of the experimental condition, indicating that whether the agent responds at fixed time points with laughter or reacts to the user's laughter does not alter how the synthesized laughter utterances and facial expressions are judged. This finding is reasonable, as the technical features, the matching of facial expression and audio features etc., were the same for all three conditions.

In more details on H1, expressing laughter in response to the users laughter behavior increased the felt amusement, as compared to an agent that acts independently of the user (i.e., fixed conditions). More specifically, they indicated that they felt the agent shared their sense of humor, the agent added to the experience in the sense that it would have been less funny without the agent, and the agent itself was funny.

Concerning the hypothesis H2 on the emotional contagion, the interactive condition rated the contagiousness of the agent's laughter as higher, compared to the fixed conditions. This confirmed the assumption, that the process of emotional contagion is facilitated by the feedback loop of the expressed laughter by the user and an adequate response by the agent.

In respect to the hypothesis H3 on the social presence, the interactive condition yielded higher scores on those evaluation items, although the difference failed to be statistically significant. Still, the numerical differences clearly favoured the interactive condition over the fixed conditions and the result may be significant if the sample size was bigger, as the cells only consisted of 9 and 11 users.

Some limitations of this study should be noted. First, in the interactive condition, the amount of laughter displayed by the agent varied for each user, depending on how many times the users actually laughed. Therefore, the agent behaved similar to the user, which seems to be natural and comfortable for the user. Nevertheless, the current state of data analysis does not allow to differentiating between individuals who displayed a lot of laughter—and consequently had a lot of laughter feedback by the agent—and individuals who showed only little laughter—and received little laughter feedback by the agent. An in–depth analysis of the video material obtained during the evaluation experiment will allow for an investigation of how many times the users actually laughed and how this influenced the perception of the setting. Moreover, the stimulus video used in the video consisted of only one type of humorous material. It is well established in psychological research that inter-individual differences exist in the appreciation of types of humor. Therefore, including only one type of humor may limit the amusement elicitation in certain users and the results may profit from a broader sample of videos. Nevertheless, for evaluating the laughing agent, this variation is not necessary. Last but not least these results need replicating in a further study with more users, as the sample size was small.

8. CONCLUSIONS AND FUTURE WORKS

In this paper we presented an interactive system that is able to detect human laughs and to respond appropriately by integrating information on the context and human behavior. We also evaluated the impact of the laugh-aware agent on the humor experience of the user and interaction.

On the technical side, the outcome of this work is a full processing chain with components that can perform multimodal data acquisition, real-time laughter-related analysis, output laughter decision and audiovisual laughter synthesis. Concerning the evaluation, the first results of the evaluation experiment are highly promising: it was shown that the conditions elicit different degrees of amusement in the user and the amount of social interaction induced, showing that the interactive condition yielded the most positive outcomes and implying that the feedback given to the user by responding to his or her laughter is best capable of creating a "mutual film watching experience" that is pleasurable.

This is ongoing work. Future works will include: improving the laughter detection and intensity estimation by using multimodal user data, extending the range of output laughs by allowing laughs to be generated or modified on the fly, as well as multimodal visual laughter synthesis including body movements and visible respiration. We also plan to evaluate other psychological measures to be able to control for influences of personality and mood on the experimental session and the evaluation of the agent. For example, gelotophobes, individuals with a fear of being laughed at [21], do not perceive any laughter as joyful or relaxing and they fear being laughed at even in ambiguous situations. Therefore, the laughing virtual agent might be interpreted as a threat and the evaluation would be biased by the individuals fear. By assessing the gelotophobic trait, individuals with at least a slight fear of being laughed at can either be excluded from further analysis, or the influence of gelotophobia can be investigated for the dependent variables.

9. ACKNOWLEDGMENTS

This work was supported by the European FP7-ICT-FET project ILHAIRE (grant $n^\circ 270780$).

10. ADDITIONAL AUTHORS

Additional authors: Huseyin Cakmak (Université de Mons) and Sathish Pammi (TELECOM ParisTech) and Tobias Baur (Universität Augsburg) and Stephane Dupont (Université de Mons) and Matthieu Geist (Supelec) and Florian Lingenfelser (Universität Augsburg) and Gary McKeown (The Queen's University of Belfast) and Olivier Pietquin (Supelec) and Willibald Ruch (Universität Zürich).

11. REFERENCES

[1] J. Urbain, R. Niewiadomski, E. Bevacqua, T. Dutoit, A. Moinet, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "AVLaughterCycle: Enabling a virtual agent to join in laughing with a conversational partner using a similarity-driven audiovisual laughter animation," *Journal on Multimodal User Interfaces*, vol. 4, no. 1, pp. 47–58, 2010.

[2] S. Shahid, E. Krahmer, M. Swerts, W. Melder, and M. Neerincx, "Exploring social and temporal dimensions of emotion induction using an adaptive affective mirror," in *Proceedings of CHI'09, Extended Abstracts Volume*. ACM, 2009, pp. 3727–3732.

[3] S. Fukushima, Y. Hashimoto, T. Nozawa, and H. Kajimoto, "Laugh enhancer using laugh track synchronized with the user's laugh motion," in *Proceedings of CHI'10*, 2010, pp. 3613–3618.

[4] C. Becker-Asano, T. Kanda, C. Ishi, and H. Ishiguro, "How about laughter? Perceived naturalness of two laughing humanoid robots," in *Affective Computing and Intelligent Interaction*, 2009, pp. 49–54.

[5] C. Becker-Asano and H. Ishiguro, "Laughter in social robotics - no laughing matter," in *Intl. Workshop on Social Intelligence Design*, 2009, pp. 287–300.

[6] J. Wagner, F. Lingenfelser, and E. André, "The social signal interpretation framework (SSI) for real time signal processing and recognition," in *Proceedings of Interspeech 2011*, 2011.

[7] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1459–1462.

[8] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent," *Trans. Affective Computing*, vol. 3, pp. 5–17, 2012.

[9] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "The AVLaughterCycle database," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010.

[10] T. Cox, "Laughter's secrets: faking it – the results," *New Scientist*, 27 July 2010.

[11] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Speech Synthesis Workshop*, Santa Monica, California, September 2002, pp. 227–230.

[12] K. Oura, "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/, consulted on June 22, 2011.

[13] J. Urbain and T. Dutoit, "A phonetic analysis of natural laughter, for use in automatic laughter processing systems," in *International Conference on Affective Computing and Intelligent Interaction*, Memphis, Tennesse, October 2011, pp. 397–406.

[14] P. Ekman, W. Friesen, and J. Hager, "Facial action coding system: A technique for the measurement of facial movement," 2002.

[15] J. Ostermann, "Face animation in MPEG-4," in *MPEG-4 Facial Animation - The Standard Implementation and Applications*, I. Pandzic and R. Forchheimer, Eds. Wiley, 2002, pp. 17–55.

[16] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[17] A. Chapman, "Humor and laughter in social interaction and some implications for humor research," in *Handbook of humor research, Vol. 1*, P. McGhee and J. Goldstein, Eds., 1983, pp. 135–157.

[18] W. Ruch, "State and trait cheerfulness and the induction of exhilaration: A FACS study," *European Psychologist*, vol. 2, pp. 328–341, 1997.

[19] A. Nijholt, "Humor and embodied conversational agents," http://doc.utwente.nl/41392/.

[20] W. Ruch and P. Ekman, "The expressive pattern of laughter," in *Emotion, qualia and consciousness*, A. Kaszniak, Ed. Tokyo: World Scientific Publishers, 2001, pp. 426–443.

[21] W. Ruch and R. Proyer, "The fear of being laughed at: Individual and group differences in gelotophobia." *Humor: International Journal of Humor Research*, vol. 21, pp. 47–67, 2008.