

Laugh When You're Winning

Maurizio Mancini, Laurent Ach, Emeline Bantegnie, Tobias Baur, Nadia Berthouze, Debajyoti Datta, Yu Ding, Stéphane Dupont, Harry J. Griffin, Florian Lingenfelser, et al.

▶ **To cite this version:**

Maurizio Mancini, Laurent Ach, Emeline Bantegnie, Tobias Baur, Nadia Berthouze, et al.. Laugh When You're Winning. 9th International Summer Workshop on Multimodal Interfaces (eNTERFACE), Jul 2013, Lisbon, Portugal. pp.50-79, 10.1007/978-3-642-55143-7_3 . hal-01350739

HAL Id: hal-01350739
https://inria.hal.science/hal-01350739

Submitted on 1 Aug 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Laugh When You're Winning

Maurizio Mancini[1], Laurent Ach[2], Emeline Bantegnie[2], Tobias Baur[3], Nadia Berthouze[4], Debajyoti Datta[5], Yu Ding[5], Stphane Dupont[6], Harry J. Griffin[4], Florian Lingenfelser[3], Radoslaw Niewiadomski[1], Catherine Pelachaud[5], Olivier Pietquin[7], Bilal Piot[7], Jérôme Urbain[6], Gualtiero Volpe[1], and Johannes Wagner[3]

[1]InfoMus - DIBRIS, Università Degli Studi di Genova, Viale Francesco Causa, 13, 16145 Genova, Italy
{maurizio.mancini,gualtiero.volpe}@unige.it, radoslaw.niewiadomski@dibris.unige.it
[2]LA CANTOCHE PRODUCTION, Hauteville, 68, 75010 Paris, France
{lach,ebantegnie}@cantoche.com
[3]Institut für Informatik, Universität Augsburg, Universitätsstr. 6a, 86159 Augsburg, Germany
{tobias.baur,florian.lingenfelser,johannes.wagner}@informatik.uni-augsburg.de
[4]UCLIC, University College London, Gower Street, London, WC1E 6BT, United Kingdom
{nadia.berthouze,harry.griffin}@ucl.ac.uk
[5]CNRS - LTCI UMR 5141 - Telecom ParisTech, 37-39 Rue Dareau, 75014 Paris, France
{debajyoti.datta,yu.ding,catherine.pelachaud}@telecom-paristech.fr
[6]TCTS Lab, Faculté Polytechnique, Université de Mons, Place du Parc 20, 7000 Mons, Belgium
{stephane.dupont,jerome.urbain}@umons.ac.be
[7]École Supérieure d'Électricité, Rue Edouard Belin, 2, 57340 Metz, France
{olivier.pietquin,Bilal.Piot}@supelec.fr

Abstract. Developing virtual characters with naturalistic game playing capabilities is an increasingly researched topic in Human-Computer Interaction. Possible roles for such characters include virtual teachers, personal care assistants, and companions for children. Laughter is an under-investigated emotional expression both in Human-Human and Human-Computer Interaction. The EU Project ILHAIRE, aims to study this phenomena and endow machines with laughter detection and synthesis capabilities. The *Laugh when you're winning* project, developed during the eNTERFACE 2013 Workshop in Lisbon, Portugal, aimed to set up and test a game scenario involving two human participants and one such virtual character. The game chosen, the yes/no game, induces natural verbal and non-verbal interaction between participants, including frequent hilarious events, e.g., one of the participants saying "yes" or "no" and so losing the game. The setup includes software platforms, developed by the ILHAIRE partners, allowing automatic analysis and fusion of human participants' multimodal data (voice, facial expression, body movements, respiration) in real-time to detect laughter. Further, virtual characters endowed with multimodal skills were synthesised in order to interact with the participants by producing laughter in a natural way.

Keywords: HCI, laughter, virtual characters, game, detection, fusion, multimodal.

1 Introduction

Computer-based characters play an ever-increasing role in Human-Computer Interaction, not only for entertainment but also for education, as assistants and potentially in healthcare. Such emotionally complex interactions demand avatars

that can detect and synthesise emotional displays. Laughter is a ubiquitous and complex but under-investigated emotional expression. The *Laugh when you're winning* eNTERFACE 2013 Workshop project builds on the work of the EU Project ILHAIRE[1] and on the previous eNTERFACE projects *AVLaughterCycle* [52] and *Laugh Machine* [33].

The project consists of an avatar actively participating in social games, in particular the *yes/no* game scenario. The avatar capabilities developed for game playing will have many applications beyond simple entertainment. The complex human-avatar interaction of a game demands considerable behavioural naturalness for the avatar to be a credible, trustworthy character. The avatar responds to user laughter in a highly customised way by producing laughter of its own.

Laughter detection and analysis among the speech, noise and body movements that occur in social games is achieved through multimodal laughter detection and analysis of audio, video, body movements and respiration. Laughter decisions integrate output from a module that drives mimicry behaviour, in response to the detected parameters of users' laughter, e.g., intensity.

The close interaction of a game scenario, proposed here, demands precise laughter detection and analysis and highly natural synthesised laughter. The social effect of avatar laughter also depends on contextual factors such as the task, verbal and nonverbal behaviours beside laughter and the user's cultural background [2, 3]. In addition social context and emotional valence have been shown to influence mimicry [5]. Therefore, in a game scenario with both positive and negative emotions, laughter and mimicry must be well-implemented in order to enhance rather than inhibit interaction.

In the last part of the report we present an experiment, carried out during the eNTERFACE 2013 Workshop, which assesses users' perception of avatar behaviour in the direct interaction involved in the game scenario. The level of emotional response displayed by the avatar is varied: no response, responsive, responsive with mimicry. Measures of users' personality are analysed alongside short-term measures, e.g., user laughter, and long-term measures of engagement, e.g., mood, trust in the avatar. This spectrum of measures tests the applicability of an emotionally sensitive avatar and how its behaviour can be optimised to appeal to the greatest number of users and avoid adverse perceptions such as a malicious, sarcastic or unnatural avatar.

2 Background

The concept of a game playing robot has long intrigued humans, with examples, albeit fake, such as the Mechanical Turk in the 18th century [38]. Games are complex social interactions and the possibility of victory or defeat can make them emotionally charged. The importance of emotional competence (the ability to detect and synthesise emotional displays) has therefore been recognised in more recent human-avatar/robot systems. Leite et al. [25] describe an empathic chess-playing robot that detected its opponent's emotional valence. More children

[1] `http://www.ilhaire.eu`

reported that the robot recognised and shared their feelings when it displayed adaptive emotional expressions during games.

Laughter often occurs in games due to their social context and energetic, exhilarating nature. Recognising and generating laughter during games is therefore vital to an avatar being an engaging, emotionally convincing game companion. In addition, a trend for gamification - "the use of game attributes to drive game-like behavior in a non-gaming context may increase emotional expressions, such as laughter, in serious or mundane tasks" [35]. Thus an emotionally competent avatar developed for a game situation may well have value in situations such as education, exercise or rehabilitation.

3 State of the Art

3.1 Laughter Installations

Previous laughter detection and response systems have generally used a limited human-avatar interaction. Fukushima et al. [15] built a system that enhanced users' laughter activity during video watching. It comprised small dolls that shook and played prerecorded laughter sounds in response to users' laughter.

AVLaughterCycle aimed to create an engaging laughter-driven interaction loop between a human and the agent [52]. The system detected and responded to human laughs in real time by recording the user's laugh and choosing an acoustically similar laugh from an audiovisual laughter database.

The Laugh Machine project endowed a virtual agent with the ability to laugh with a user as a fellow audience member watching a funny video [53, 33]. The agent was capable of detecting the participant's laughs and laughing in response to both the detected behaviour or to pre-annotated humorous content of the stimulus movie. The system was evaluated by 21 participants taking part in one of three conditions: interactive laughter (agent reacting to both the participant's laughs and the humorous movie), fixed laughter (agent laughing at predefined punchlines of the movie) or fixed speech (agent expressing verbal appreciation at predefined punchlines of the movie). The results showed that the interactive agent led to increased amusement and felt contagion.

3.2 Laughter Detection

Laughter has long been recognised as a whole-body phenomenon which produces distinctive body movements. Historical descriptions of these movements include bending of the trunk, movement of the head and clutching or slapping of the abdomen or legs [40]. The distinctive patterns of respiration that give rise to the equally distinctive vocalisations of laughter also generate movements of the trunk. An initial rapid exhalation dramatically collapses the thorax and abdomen and may be followed by a series of smaller periodic movements at lower volume. Fukushima et al. used EMG signals reflecting diapragmatic activity involved in this process to detect laughter [15]. These fundamental laughter

actions also drive periodic motion elsewhere in the body. Motion descriptors based on energy estimates, correlation of shoulder movements and periodicity to characterise laughter have been investigated [29]. Using a combination of these measures a Body Laughter Index (BLI) was calculated. The BLIs of 8 laughter clips were compared with 8 observers' ratings of the energy of the shoulder movement. A correlation, albeit weak, between the observers' ratings and BLIs was found. This model is used in the current project (see Section 6.2).

A body of work on recognition of emotion from body movements has accumulated in recent years [20, 21, 9, 4, 30]. Some of this work has concentrated on differences in movements while walking. Analysing the body movements of laughter presents a contrasting challenge in that, unlike walking, its emotional content cannot be modelled as variations in a repeated, cyclical pattern. Furthermore, body movements related to laughter are very idiosyncratic. Perhaps because of this, relatively little detection of laughter from body movements (as opposed to facial expressions) has been undertaken. Scherer et al. [43] applied various methods for multimodal recognition using audio and upper body movements (including head). Multimodal approaches actually yielded less accurate results than combining two types of features from the audio stream alone. In light of these results there is obviously considerable room for improvement in the contribution of body-movement analysis to laughter detection.

Discrimination between laughter and other events (e.g., speech, silence) has for a long time focused only on the audio modality. Classification typically relies on Gaussian Mixture Models (GMMs) [47], Support Vector Machines (SVMs) [47, 19], Multi-Layer Perceptrons (MLPs) [22] or Hidden-Markov Models (HMMs) [8], trained with traditional spectral and prosodic features (MFCCs, PLP, pitch, energy, etc.). Error rates vary between 2 and 15% depending on the data used and classification schemes. Starting from 2008, Petridis and Pantic enriched the so far mainly audio-based work in laughter detection by consulting audio-visual cues for decision level fusion approaches [36]. They combined spectral and prosodic features from the audio modality with head movement and facial expressions from the video channel. They reported a classification accuracy of 74.7% in distinguishing three classes: unvoiced laughter, voiced laughter and speech [37]. Apart from this work, there exists to our knowledge no automatic method for characterizing laughter properties (e.g., emotional type, arousal, voiced or not). It must also be noted that few studies have investigated the segmentation of continuous streams (as opposed to classifying pre-segmented episodes of laughter or speech) and that performance in segmentation is lower than classification performance [37].

3.3 Laughter Acoustic Synthesis

Until recently, work on the acoustic synthesis of laughter has been sparse and of limited success with low perceived naturalness. We can for example cite the interesting approach taken by Sundaram an Narayanan [44], who modeled the rhythmic energy envelope of the laughter acoustic energy with a mass-spring model. A second approach was the comparison of articulatory synthesis and

diphone concatenation done by Lasarcyk and Trouvain [24]. In both cases the synthesizd laughs were perceived as significantly less natural than human laughs. Recently, HMM-based synthesis, which had been efficiently applied to speech synthesis [46], has advanced the state-of-the-art [49].

3.4 Laughter Synthesis with Agents

Few visual laughter synthesis models have been proposed so far. The major one is by Di Lorenzo et al. [13] who proposed an anatomically inspired model of upper body animation during laughter. It allows for automatic animation generation of the upper-body from a preregistered sound of laughter. Unfortunately it does not synthesize head and facial movement during laugh. Conversely, a model proposed by Cosker and Edge [11] is limited to only facial animation. They built a data-driven model for non-speech related articulations such as laughs, cries etc. It uses HMM trained from motion capture data and audio segments. For this purpose, the number of facial parameters acquired with and optical motion capture system Qualisys was reduced using PCA, while MFCC was used for the audio input. More recently Niewiadomski and Pelachaud [32] have proposed a model able to modulate the perceived intensity of laughter facial expressions. For this purpose, they first analysed the motion capture data of 250 laughter episodes annotated with 5-point intensity scale and then extracted a set of facial features that re correlated with the perceived laughter intensity. By controlling these features the model modulates the intensity of displayed laughter episodes.

4 Innovation: Multimodality

As already explained, the *Laugh When You're Winning* project builds upon the Laugh Machine project that was carried out during eNTERFACE'12. The major innovations with regards to Laugh Machine or other installations are the following (these innovations will be further detailed in Sections 6 and 7):

- The laughter detection module has been extended to take multimodal decisions: estimations of the likelihoods of Smile, Speaking and Laughter likelihoods result from analyses of audio, facial and body movements, while in Laugh Machine there was simply a laughter/no-laughter detection based on audio only; in addition, the intensity estimation module has been improved (a neural network was trained under Weka).
- Several modules exchange information in real time for detection and analysis of laughter: the master process is Social Signal Interpretation (SSI) but some computations are outsourced to Eyesweb and Weka.
- The new game scenario, which can foster different types of emotions and involves 2 users simultaneously taken into account by the system; an ad hoc game engine has been developed to manage this specific scenario.
- The integration of laughter mimicry, through modules that analyse some laughter properties of (one of) the participants (e.g., shoulder movements periodicity) to influence the laughs displayed by the agent (shoulder periodicity and rhythm of the acoustic signal are driven by the measured properties).

5 Social Signal Interpretation (SSI)

The recognition component has to be equipped with certain sensors to capture multimodal signals. First, the raw sensor data is collected, synchronized and buffered for further processing. Then the individual streams are filtered, e.g. to remove noise, and transformed into a compact representation by extracting a set of feature values from the time- and frequency space. In this way the parameterized signal can be classified by either comparing it to some threshold or applying a more sophisticated classification scheme. The latter usually requires a training phase where the classifier is tuned using pre-annotated sample data. The collection of training data is thus another task of the recognition component. Often, activity detection is required in the first place in order to identify interesting segments, which are subject to a deeper analysis. Finally, a meaningful interpretation of the detected events is only possible in the context of past events and events from other modalities. For instance, detecting several laughter events within a short time frame increases the probability that the user is in fact laughing. On the other hand, if we detect that the user is talking right now we would decrease the confidence for a detected smile. The different tasks the recognition component undertakes are visualized in Figure 1.

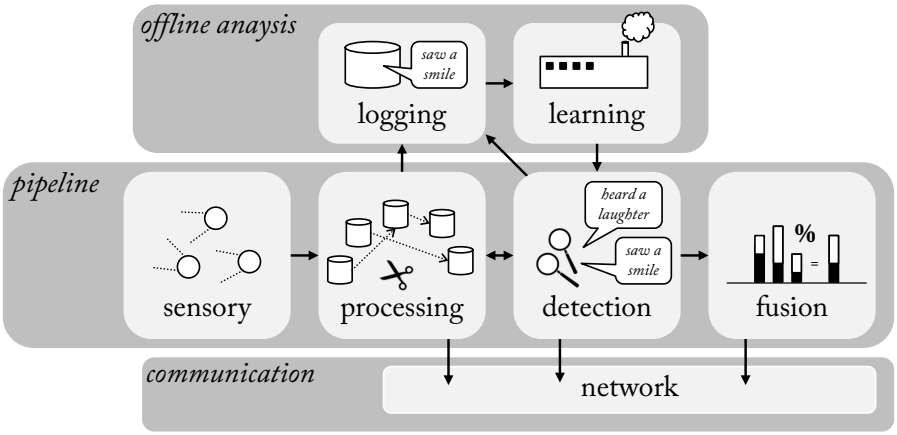


Fig. 1. Scheme of the laughter recognition component implemented with the Social Signal Interpretation (SSI) framework. Its central part consists of a recognition pipeline that processes the raw sensory input in real-time. If an interesting event is detected it is classified and fused with previous events and those of other modalities. The final decision can be shared through the network with external components.

The Social Signal Interpretation (SSI) software [54] developed at Augsburg University suits all mentioned tasks and was therefore used as a general framework to implement the recognition component. SSI provides wrappers for a large

range of commercial sensors, such as web/dv cameras and multi-channel ASIO audio devices, or the Microsoft Kinect, but other sensors can be easily plugged to the ystem thanks to a patch-based architecture. It also contains processing modules to filter and/or extract features from the recording signals. In addition, it includes several classifiers (K-nearest Neighbor, Support Vector Machines, Hidden Markov Models, etc.) and fusion capabilities to take unified decisions from several channels.

In this project, SSI was used to synchronize the data acqusition from all the involved sensors and computers, estimate users' states (laughing, speaking or smiling) from audio (see Laugh Machine project [33]) and face (see Section 6.1, as well as fusing the estimations of users' states coming from the different modalities: audio, face anlysis and body analysis ((outsourced to Eyesweb, see Section 6.2).

6 Multimodal Laughter Detection

6.1 Face Analysis

Face tracking provided by the Kinect SDK gives values for 6 action units (AUs) that are used to derive the probability that the user is smiling (in particular position of the upper lip and lip corners). In our tests we selected 4 of them as promising candidates for smile detection, namely *upper lip raiser*, *jaw lowerer*, *lip stretcher*, *lip corner depressor* (see Figure 2). In order to evaluate these features test recordings were observed and analysed in Matlab.



Fig. 2. Promising action units for smile detection provided by Kinect face tracking, namely *upper lip raiser*, *jaw lowerer*, *lip stretcher*, *lip corner depressor*

Plots of the features over time are visualized in Figure 3. Laughter periods are highlighted in green. We can see that especially the values received for *upper lip raiser* (1st graph) and *lip stretcher* (3rd graph) are significantly higher during laughter periods than in-between laughter periods; *lip corner depressor*, on the other hand, has a negative correlation, i. e. values decrease during periods of laughter.

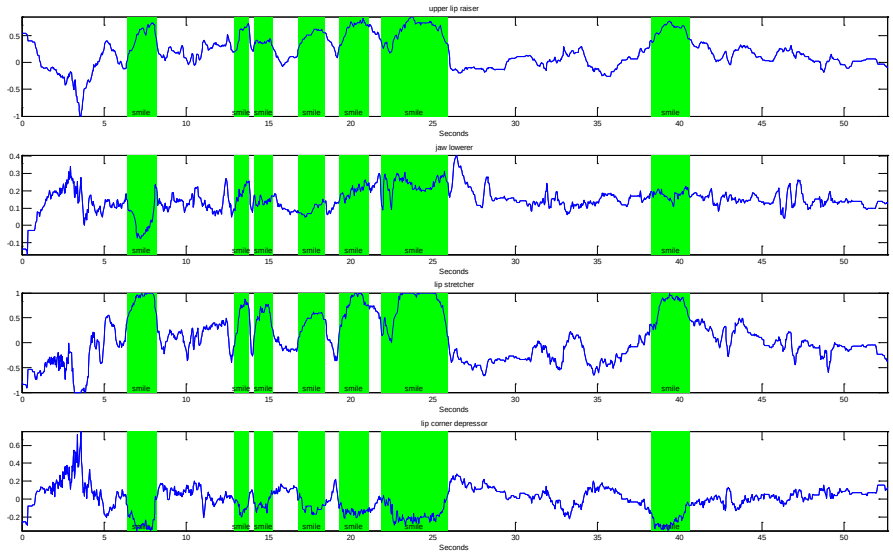


Fig. 3. Correlation between the measured action units and periods of laughter (green)

In order to combine the action units to a single value we found the following formula to give reasonable good results:

$$p_{smile} = \textit{upper lip raiser} \times \textit{lip stretcher} \times (1 - \textit{lip corner depressor}) \quad (1)$$

In order to filter out natural trembling we additionally define a threshold T_{smile}. Only if above the threshold, p_{smile} will be included in the fusion process (see Section 6.3). In our test $T_{smile} = 0.5$ gave good results.

As a second source of information Fraunhofer's tool SHORE [42] is used to derive a happy score from the currently detected face. Tests have shown that the happiness score highly correlates with user smiles. Including both decisions improves overall robustness.

6.2 Body Analysis

Real-time processing of body (i.e., trunk, shoulders) features is performed by EyesWeb XMI [27]. Compressed (JPEG) Kinect depth image streams captured by SSI are sent on-the-fly via UDP packets to a separate machine on which EyesWeb XMI programs (called *patches*) detect shoulder movements and other body-movement measures, e.g., Contraction Index. Additionally, color-based tracking of markers (green polystyrene balls) placed on the user's shoulders is performed by EyesWeb XMI and the resulting recognition data is sent back to SSI to be integrated in the following overall laughter detection and fusion process (Section 6.3).

The body detection algorithms we present in this report are an improvement and extension of the techniques developed for the Laugh Machine (eNTERFACE'12) [33]. In particular, the previously described Body Laughter Index (BLI) is computed as a weighted sum of user's shoulders correlation and energy:

$$BLI = \alpha\bar{\rho} + \beta\bar{E} \tag{2}$$

where the correlation ρ is computed as the Pearson correlation coefficient between the vertical position of the user's left shoulder and the vertical position of the user's right shoulder; and kinetic energy E is computed from the speed of user's shoulders and their mass relative to body mass.

We also validate the BLI by the user's shoulder movement frequency: if frequency is included in an acceptable interval $[2, 8]Hz$ then the BLI is valid. The interval is motivated by psychological studies on laughter by Ruch and Ekman [40].

In this report we introduce a new information for the body (i.e., trunk, shoulders) modality: laughter intensity. When a laughter event is detected by using the BLI, the FFT of the Y component of shoulders and trunk is computed along the entire event length (events lasted from 1 second to 9 seconds). The two most prominent peaks of the FFT, $max1$ (the absolute maximum) and $max2$ (the second most prominent peak) are then extracted. These are used to compute the following index:

$$R = \frac{max1 - max2}{max1} \tag{3}$$

Basically, the index will tend to 1 if just one prominent component is present; it will tend to 0 if two or more prominent components are present. Thus, periodic movements, i.e., those exhibiting one prominent component, will be characterized by and index near 1, while the index for non-periodic movements will be near 0. Figure 4 shows two examples of such computation: on the left, one peak around 1/3 Hz is shown, probably related to torso rocking during laughter, and the index tends is close to 1, indicating a highly periodic movement; on the right, many peaks between 1/3 Hz and 1.5 Hz are shown, and the index is close to 0, indicating a mildly periodic movement.

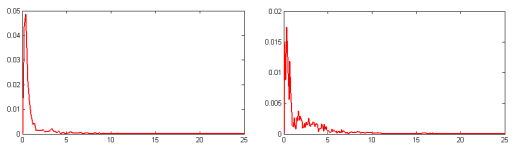


Fig. 4. FFT computed on the shoulder Y coordinate. On the left a prominent component is present and the index tends to 1. On the right many prominent components are present and the index tends to 0.

A test carried out in the past months on 25 laughter events annotated for intensity by psychologists [28], showed that R can successfully approximate laughter intensity. Significant correlations between R and the manually annotated intensity values were found for both shoulders ($r = 0.42, p = 0.036$) and trunk ($r = 0.55, p = 0.004$).

Table 1. Correlation between body indexes and annotated laughter intensity

Index	Correlation	p-Value
Rs	0,4216	0,0358
Rt	0,5549	0,0040
Rd	0,1860	0,3732

Table 1 reports correlation and p-values for shoulder/trunk indexes and annotated laughter intensity. Rs is the index computed only on shoulder movement; *Rt* is the same index computed only on trunk movement; *Rd* is the index computed on the difference between shoulder and trunk movement (that is, shoulder movement relative to trunk position).

6.3 Detection and Fusion

During fusion a newly developed event based vector fusion enhances the decision from the audio detector (see [33]) with information from the mentioned sources. Since the strength of the vectors decays over time, their influence on the fusion process decreases, while they still contribute to keep recognition stable. The final outcome consists of three values expressing probability for talking, laughing and smiling.

The method is inspired by work by Gilroy et al. [16] and adapted to the general problem of fusing events in a single or multi-dimensional space. A main difference from their proposed method is that a modality is not represented by a single vector, but new vectors are generated with every new event. In preliminary experiments this led to much more stable behaviour of the fusion vector, since the influence of false detections is lowered considerably. Only if several successive events are pointing in the same direction is the fusion vector dragged into this direction.

The algorithm is illustrated in Figure 5. In the example three successive events are measured: a cough sound, a shoulder movement shortly after that and, after a small delay, a smile. Each event changes the probability that the user is laughing. When the first event, the cough sounds, arrives it is still unlikely, since, although it is a vocal production, coughing differs from laughter. However, a shoulder movement is detected shortly after, laughter becomes more likely and the laughter probability is increased. And when finally a smile is detected the laughter probability becomes even more likely. Due to the decay function that is applied to the event vectors the probability afterwards decreases over time.

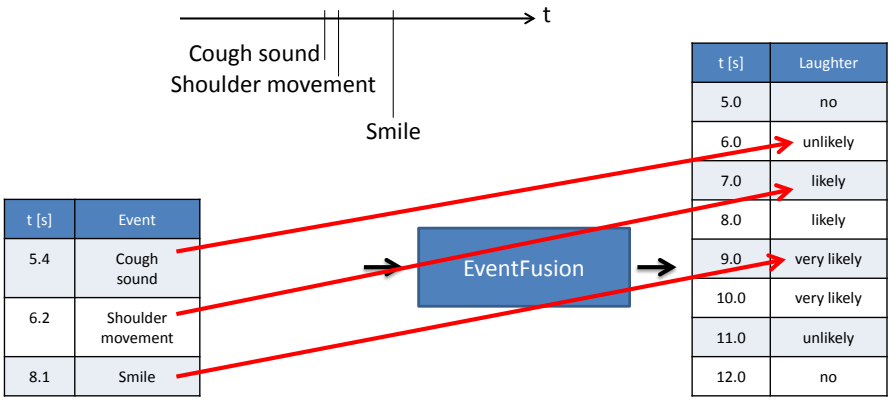


Fig. 5. Example of event fusion

Thanks to the fusion, performance in terms of reliability and robustness has clearly been improved compared to the previous system. A schema of the final detection system is shown in Figure 6.

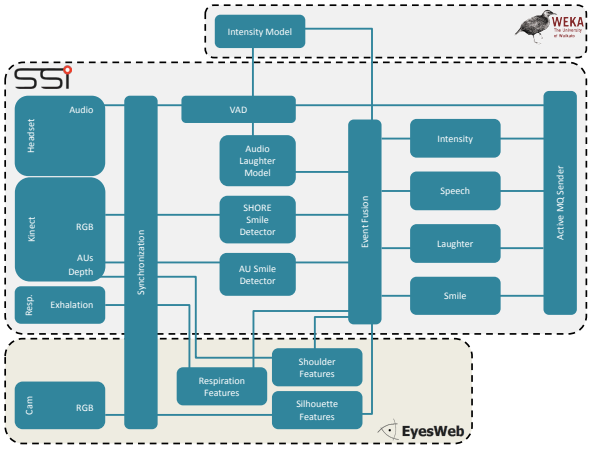


Fig. 6. The final detection system

7 Multimodal Laughter Synthesis

7.1 Dialogue Manager

The original objective of the project was to train a dialogue manager from human data; however, this component could not be built within the time constraints

of the project. To allow for the interaction to take place, a rule-based dialogue manager with empirical thresholds was designed. It follows simple rules to decide when and how (duration, intensity) the agent should laugh, given the state of the game (game in progress, game finished) and the detected states of the two participants (speaking, laughing, smiling or none of these). The implemented rules are presented in Table 2. Empirical thresholding on the speaking, laughing and smiling likelihoods was used to determine the state of each participant. The implemented rules are symmetric with respect to the two participants (no difference is made between speaker and observer, the same rules apply if the participants are switched).

Table 2. Rules for determining when and how the agent should laugh. The implemented rules are symmetric (Participant1 and Participant2 can be reversed). If several rules are met (e.g. likelihoods for Laughter and Speech of Participant 1 both reach the defined thresholds, the highest rule in the table receives priority.

Participants states		Laughter decision		Participants states		Laughter decision	
P1	P2	Intensity	Duration	P1	P2	Intensity	Duration
Laugh	Laugh	High	High	Speak	Smile	Low	Low
Laugh	Speak	Low	Low	Speak	Silent	/	/
Laugh	Smile	Medium	Medium	Smile	Smile	Medium	Medium
Laugh	Silent	Medium	Medium	Smile	Silent	Low	Low
Speak	Speak	/	/	Silent	Silent	/	/

The dialog manager also considers the game context, which is obtained thanks to mouse clicks send by SSI. A click on Mouse 1 signals the start of a yes/no game. A click on Mouse 2 tells that the speaker has lost the game (by saying "yes" or "no"). Thanks to these clicks, the dialog manager can determine at every moment the state of the game, which can take 4 different values: game not started yet, game on, game lost, game won[2]. This information on the game state is further transmitted to the laughter planner.

7.2 Laughter Planner

Laughter Planner controls the behavior of the agent as well as the flow of the game. It decides both verbal and nonverbal messages taking into account the verbal and nonverbal behavior of the human participants of the game, who are denoted the speaker (SPR; the person that is challenged in the game) and the observer (OBS; i.e. the second human player that also poses the questions), and the rules of the game. Laughter Planner receives continuously the inputs presented in Table 3.

[2] In our case, the game is won if the speaker manages to avoid saying yes or no during 1 minute, so the dialog manager puts the game status to game won one minute after the game started (click on Mouse 1), if there was no click on Mouse 2 (game lost) in the meantime

Table 3. Laughter Planner Inputs (DM = Dialog Manager; MM = Mimicry Module)

Name	Description	Values	Sender
LAUGH_DUR	Duration of the laugh to be displayed by the agent	R+	DM
LAUGH_INT	Intensity of the laugh to be displayed by the agent	[0, 1]	DM
MIMICKED_AMP	relative amplitude of human laughter	[-1, 1]	MM
MIMICKED_VEL	relative velocity of human laughter	[-1, 1]	MM
SPEECH_P_SPR	probability that the speaker is currently speaking	[0, 1]	SSI
SPEECH_P_OBS	probability that the observer is currently speaking	[0, 1]	SSI

The main task of Laughter Planner is to control the agent behavior. The details of the algorithm are presented in Figure 7. Laughter Planner generates both the questions to be posed by the agent to the human player as well as laughter responses.

The game context is also taken into account: the agent is only allowed to ask questions when the game is on; when the game is won, the agentinforms the participants (e.g., "Congratulations, the minute is over, you won!"); when the game is lost, the agent laughs in the laughter conditions, or says something in the no-laughter condition e.g., "Oh no, you just lost").

The questions are selected from the pre-scripted list of questions. This list contains the questions that were often used by humans when playing the game (e.g. MMLI corpus [31]). Some of the questions are independent of others while others are asked only as part of a sequence e.g. "What's your name?"... "Is that your full name"... "Are you sure?". The questions in sequence should be displayed in the predefined order, while the order of other questions is not important. Additionally, Laughter Planner takes care not to repeat the same question twice in one game. Each question is coded in BML that is sent at the appropriate moment to any of two available Realizers (Greta, Living Actor). The Planner poses a new question when neither of the human participants speak for at least 5 seconds. If the observer starts to speak, he probably poses a new question or a new sequence of questions. In that case, Laughter planner abandons its sequence of questions and starts a new one in the next turn.

Also the set of laughter is predefined. The laughte episodes (audio and facial expressions) are pre-synthesized off-line (see Sections 7.4 and 7.5 for details) from the available data (AVLC corpus). Only the shoulder movements are generated in real time. For each episode of AVLC corpus, five different versions were created, each of them with different laugh burst duration and consequently also different durations. Thus each original sample can be played "quicker" or "slower" and also corresponding lip movement animation and shoulder movement animation are accordingly modified. All the pre-synthesized laughs are divided into 3 clusters according to their duration and intensity. Additionally each cluster is divided into 5 subclusters according to the mean laugh burst velocity. While the choice of the episode is controlled with 2 input parameters sent by Dialog Manager (see Table 3), the 2 parameters sent by Mimicry Module are used to choose the correct velocity variation of the episode. In more details,

the values of LAUGH_DUR and LAUGH_INT are used to choose a cluster of laugh episodes. Next, mimicry parameters are used to choose a subcluster of this cluster of episodes, i.e. a set of laughs of laugh bursts corresponding to the value sent by mimicry module. Finally, the laugh episode is chosen randomly from the subcluster and BML messages containing the name of episode as well as BML tags describing the animation over different modalities are sent to the Realizer.

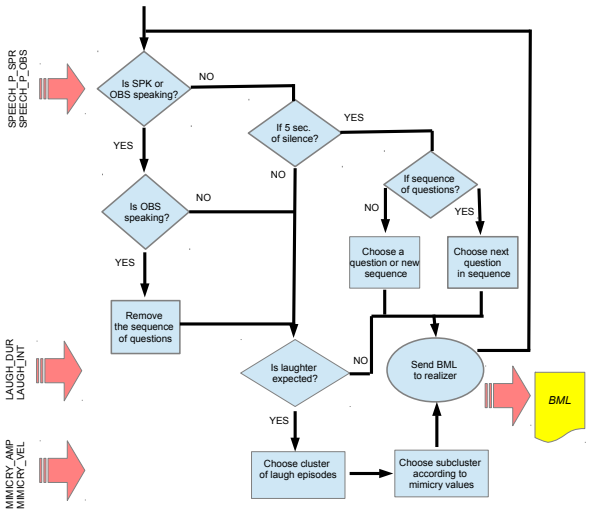


Fig. 7. Laughter Planner

7.3 Mimicry

The mimicry module has the task of deciding how the agent's expressive behaviour should mimic the user's one. The Greta agent has the capability to modulate its quality of movement (e.g., amplitude, speed, fluidity, energy, etc) depending on a set of *expressivity parameters*.

As illustrated in Figure 8, the mimicry module receives a vector of the user's body movement features X (see Section 6.2) as well as laughter probability (FLP) and intensity (FLI) resulting from the modality fusion process (see Section 6.3).

The mimicry module starts to work in *non-laugh* state. When FLP passes a fixed threshold T_1, the mimicry module enters the *laugh* state and starts to accumulate body features in X_E. In order to avoid continuous fast switching between laugh and non-laugh state, FLP is then compared against a second, lower, threshold T_2. When FLP goes under this threshold the mimicry module goes back to the non-laugh state. This means that the laughter event ends and a few operations are performed:

– the vector of the laughter event mean body features is computed as the ratio between X_E and the duration, in frames, of the event $count_E$;
– the duration of the event, in seconds, d_E is computed;
– the *overall* mean body features vector X_A is computed as the incremental mean of X_E;
– the *overall* mean event duration d_A is computed as the incremental mean of d_E;
– the mean body features vector X_E and the event duration d_E are stored into a file for later offline use;

Finally, the overall mean body features vector X_A and event duration d_A are sent to the Laughter Planner (see Section 7.2) where they will contribute to modulate the agent's laughter duration and body expressive features.

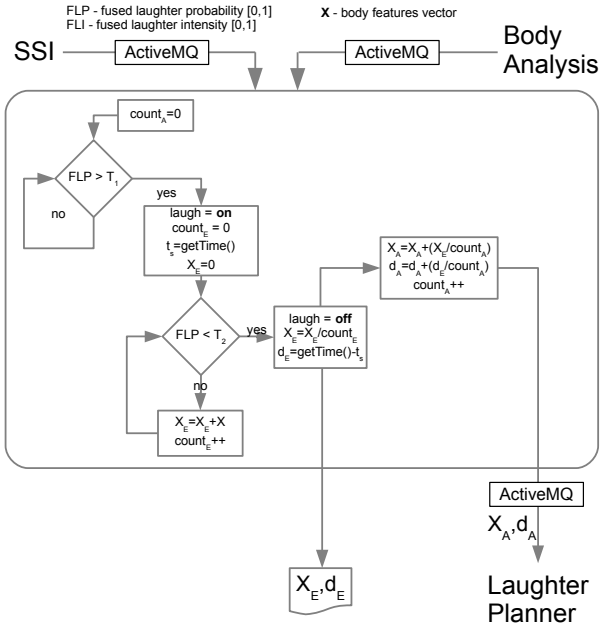


Fig. 8. Mimicry module.

7.4 Acoustic Laughter Synthesis

Acoustic laughter synthesis technology is the same as presented in [50]. It relies on Hidden Markov Models (HMMs) trained under HTS [34] on 54 laughs uttered by one female participant of the AVLaughterCycle recordings [52]. After building the models, the same 54 laughs have been synthesized using as only input to the phonetic transcriptions of the laughs. The best 29 examples have been selected for the current experiments (the other 25 examples had disturbing

or badly rendered phones due to limited number of the corresponding phones in the training data).

To increase the number of available laughs and the reactivity of the system, phonetic transcriptions of laughter episodes composed of several bouts (i.e., exhalation parts separated by inhalations) have been split into bouts by cutting the original transcription at the boundaries between inhalation and exhalation phases. This indeed increases the number of laughter examples (for example one episode composed of three bouts will produce three laughter segments instead of one). This method also increases reactivity of the system - which is limited by the impossibility of interrupting currently playing laughs - as shorter laughter segments are available: instead of the original episode of, for example, 15s, the repository of available laughter now includes three bouts of, for example, 6, 4 and 5s, which would "freeze" the application for a shorter time than the initial 15s.

To enable mimicry of the rhythm in the application, several versions of the laughs have been created: laughs are made rhythmically faster or slower by multiplying the durations of all the phones in the laughter phonetic transcription by a constant factor F. The laughs corresponding to the modified phonetic transcriptions are then synthesized through HTS, with the duration imposed to respect the duration of the phonetic transcription (in other words, the duration models of HTS are not used). Laughs have been created with this process for the following values of F: 0.6, 0.7, 0.8, 0.9, 1 (original phonetic transcription), 1.1, 1.2, 1.3 and 1.4.

Finally, the acoustically synthesized laughs are placed in the repository of available laughs, which contains for each laugh: a) the global intensity of the laugh, derived from the continuous intensity curve computed as explained in [48]; b) the duration of the laugh; c) the audio file (.wav); d) the phonetic transcription of the laughs, including the intensity value of each phone; e) the rhythm of the laugh, computed as the average duration of "fricative-vowel" or "silence-vowel" exhalation syllables of the laugh.

The first two pieces of information are used for selecting the laugh to play (using the clustering process presnted in section 7.2). The next two (audio and transcription files) are needed by the agent to play the selected laugh. Finally, the rhythm of the laugh is used to refine the selection when mimicry is active (only laughs within a target rhythm interval are eligible at each moment).

7.5 Greta

Facial Animation

As for the audio signal (Section 7.4), our work is based on the AVLC data set [52]. 24 subjects (9 women and 15 men) were recorded while watching humorous videos. This corpus includes 995 laughter examples: video, audio and facial motion capture data. Laughs were phonetically annotated [51]. Automatic landmark localization algorithm was applied to all the laughter example videos for extracting the trajectories of Facial Animation Parameters (FAPs) (see [39]).

In our model we use 22 lip FAPs as lip motion features, 3 head rotation FAPs as head features and 8 eyebrow FAPs as eyebrow features. Therefore, we have the lip, head and eyebrow motions and phonetic information of all the laughter examples included in AVLC.

Lip movements play an important role in human voice production. They are highly synchronized with spoken text, e.g., phoneme. Humans can easily perceive whether spoken text and visual lip motion are synchronized. Therefore, virtual agents should be capable of automatically generating believable lip motion during voice production. Phonetic sequences have been used to synthesize lip movements during speech in previous papers [6, 10, 7, 23, 14, 12, 26], most of which use the mapping between lip form (visual viseme) and spoken phoneme. To our knowledge, no effort has focused on natural synthesis of laughter lip motions.

One of our aims is to build a module that is able to automatically produce lip motion from phonetic transcriptions (i.e., a sequence of laughter phones, as used for acoustic synthesis). This work is based on the hypothesis that there exists a close relationship between laughter phone and lip shape. This relationship is learned by a statistical framework in our work. Then the learnt statistical framework is used to synthesize the lip motion from pseudo-phonemes and duration sequences.

We used a Gaussian Mixture Model (GMM) to learn the relationship between phones and lip motion based on the data set (AVLC). The trained GMM is capable of synthesizing lip motion from phonetic sequences. One Gaussian distribution function was learnt to model the lip movements for each of the 14 phonetic clusters used for laughter synthesis. Therefore, the trained GMM was comprised of 14 Gaussian distribution functions. For synthesis, one phonetic sequence including the duration of each phone is taken as the input, which is used to establish a sequence of Gaussian distribution functions. The determined sequence of Gaussian distribution functions [45] is used to synthesize directly the smoothed trajectories. Figure 9 shows an example of synthesized lip motion.



Fig. 9. Lip motion synthesized from a phonetic transcription

Head and eyebrow behaviours also play an important role in human communication. They are considered as auxiliary functions of speech for completing the human expressions. For example, they can convey emotional states and intentions. Humans are skilled in reading subtle emotion information from head and eyebrow behaviours. So, human-like head and eyebrow behaviour synthesis is necessary for a believable virtual agent. In consequence, we wanted to synthesize head and eyebrow motion in real time from the phonetic sequences. The proposed approach is based on real human motions recorded in the database. All

the motion data sequences in the database were segmented according to the annotated phonetic labels. The motion segments were categorized into 14 clusters corresponding to the 14 phonetic classes.

We developed a new algorithm for selecting an optimal motion segment sequence from the 14 motion segment clusters, according to the given phonetic sequence. In the proposed algorithm, one cost function is defined to evaluate the costs of all the motion segments belonging to the cluster corresponding to the given phonetic label. The cost value consists of two sub-cost functions. The first sub-cost called duration cost is the difference between the motion segment duration and the target duration; the second sub-cost called position cost is the position distance between the value of the first frame of the motion segment and the value of last frame of the previously selected motion segment. The motion segment with the smallest cost value is selected.

Shoulder Movement

Previously the analysis of the motion capture data of the Multimodal Multiperson Corpus of Laughter in Interaction (MMLI) [31] has shown regularities in the shoulder movements during the laughter. In more detail, 2D coordinates of the shoulders' positions were processed using the Fast Fourier Transform (FFT). The results showed peaks in the frequency range $[3, 6]Hz$. Interestingly, from the analysis of acoustic parameters we know that similar frequencies were observed in audio laugh bursts [1, 40]. Both these sources of information were used to generate shoulder movements that are synchronized with the synthesised audio (see Section 7.4).

The shoulder movements in the Greta agent are controlled by BML tags sent by Laughter Planner. The tag *shoulder* specifies the duration of the movement as well as its two additional characteristics: period and amplitude. These parameters are chosen by the Laughter Planner (see Section 7.2). In particular the period of the movement corresponds to the mean duration of the laugh burst in the laughter episode to be displayed. The amplitude of the shoulder movement corresponds to the amplitude of the movements detected within the Mimicry Module. If the detected movements are large then also the amplitude of the agent movements is higher, and conversely. Next, the shoulders' BML tags with all these parameters are turned into a set of frames. The vertical position of the shoulder joints is computed for each frame by using the following function:

$$X(t) = Amplitude * cos(2 * PI * frequency * t - 75.75) \tag{4}$$

where the *amplitude* and *frequency* are parameters of the BML.

7.6 Living Actor™

The Living Actor™ module includes a 3D real-time rendering component using Living Actor™ technology and a communication component that constitutes the interface between the Living Actor™ avatar and the ActiveMQ messaging

system. This version is based on sample animations created by 3D artists and combines "laughter" faces (facial expressions associated with visemes that are mouth movements corresponding to synthesized laughter sounds), "laughter" body animations corresponding to several types of movements (backward bending, forward bending, shoulder rotation) and "laughter" intensities. The main animation component is related to the trunk and arms that are combined with additional animations of head and shoulders.

The prepared trunk animations are later connected to form a graph so the avatar changes its key body position (State) using transition animations. The states in the graph (see Fig. 10) correspond to different types of laughing attitudes (bending forward, bending backward, shoulder rotations). Head and shoulder back-and-forth movements are not part of this graph; they are combined with graph transitions at run time. Some low amplitude animations of the arms are added to trunk animations so the avatar does not look too rigid.

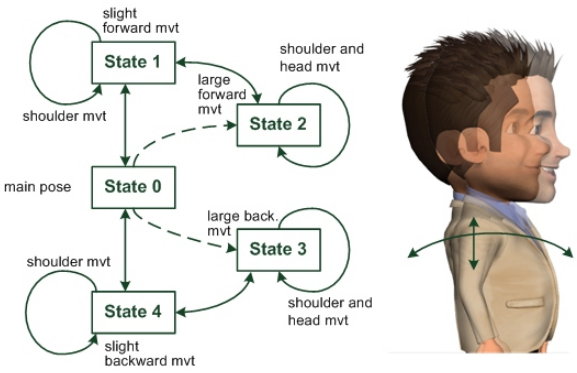


Fig. 10. Sample laughter graph of animation

Living Actor^{TM} software is originally based on graphs of animations that are combined with facial expressions and lips movements. Two main capabilities have been added to this mechanism:

- combine several animations of the body (torso, head, shoulder)
- use special facial expressions corresponding to laughter phones

The software is now able to receive data about phones and laughter intensity in real time. Depending on the received laughter intensity, a target state is chosen in the graph and transitions are followed along a path computed in real time. The input data, that include specific types of "laughter" movements, like bending forward or backward, are taken into account to choose the target states. Otherwise, one of the available types of movements is chosen by the avatar module, depending on intensity and random parameters.

The animations triggered by the graph traversal are combined with head and shoulders back-and-forth movements that make the avatar "laughter" an-

imations more realistic and avoid the perception of repetition when the same state is targeted several times in the graph. The data received from synthesized phonemes in real time are used to add facial morphing and lips movements.

When there is no instruction, the 3D real-time rendering component automatically triggers "Idle animations, so the avatar breathes, glances, or moves slightly and is never static.

8 Experiment

A preliminary experiment was run with the aim of evaluating the integrated architecture and the effect of the mimicry model on the participants. The avatar Greta was used for this experiment.

Eighteen participants (12 male, average age 26.8 (3.5) - 5 participants did not report their age) from the eNTERFACE workshop were recruited. They were asked to play the Yes/No game with the avatar in pairs. In the game, participants take turns in asking questions (observer) with the aim of inducing the other participant (speaker) to answer "yes" or "no". Each turn lasted a maximum of 1 minute or until the participant answering the questions said "yes" or "no". The avatar always played the role of supporting the observer by asking questions when a long silence occurred.

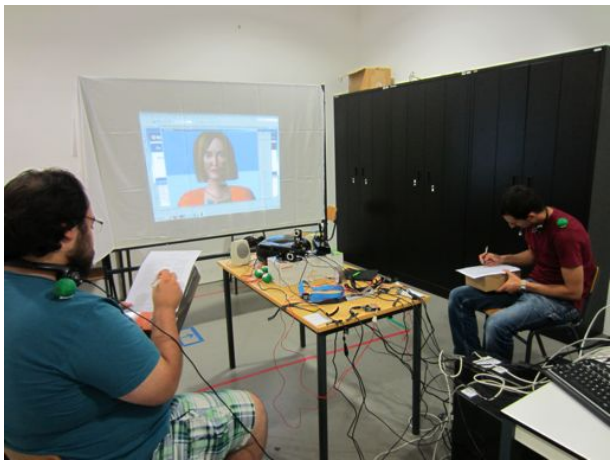


Fig. 11. Setting of the experiment. The participants are filling in an in-session questionnaire.

A within-subjects design was used: participants were asked to play the game in three different conditions: avatar talking but without exhibiting any laughter expression (No-Laughter condition), avatar exhibiting laughter expressions (Laughter condition), avatar with laughter expression and long term mimicry

capabilities (Mimicry condition). In all three conditions the avatar had laughter detection capabilities. In both the Laughter and the Mimicry conditions, the laughter responses were triggered by the detection of the laughter or smile in at least one of the participants (see Section 7.1). The order of the conditions was randomized. Each condition involved two turns of questioning, one for each participant.

The setting of the experiment is shown in Figure 11. The participants and the avatar sat around a table as shown in the figure. Each participant was monitored by a Microsoft Kinect and two webcams placed on the table. They were also asked to wear a custom made respiration sensor around their chest and a microphone around their neck.

Before the experiment, the participants had the game explained to them and were asked to sign a consent form. They were also asked to fill in a set of pre-experiment questionnaires:

- "PhoPhiKat-45": this provides scales to quantify levels of gelotophobia (the fear of being laughed at), gelotophilia (the joy of being laughed at), and katagelaticism (the joy of laughing at others) [41]. Questions are answered on a 4-point scale (1-4) and a person is deemed to have a slight expression of gelotophobia if their mean score is above 2.5 and pronounced gelotophobia if ther mean score is greater than 3.
- A Ten Item Personality Inventory (TIPI): this measure is a 10-item questionnaire used to measure the five factor personality model commonly known as the "big five" personality dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism [17].
- Current general mood: cheerfulness, seriousness and bad mood rated on a 4-point scale.
- Avatar general perception: this questionnaire measures the level of familiarity with, and the general likeability and perceived capability of avatars through a 8-item questionnaire.

After each condition, the participants were asked to fill in a questionnaire to rate their experience with the avatar (in-session questionnaire [18]). . This questionnaire is a revised version of the LAIEF-R questionnaire developed for the evaluation experiment run at eNTERFACE'12. The new version includes questions about mimicry and body expression perception and is hereafter called LAIEF-Game.

At the end of the experiment, the participants were also asked to provide comments about the overall system [18]. Each experiment lasted about 1 hour.

A second round of four games was then played in one of the two remaining conditions (randomly assigned), followed by the same questionnaire answering. Then, a last round of four games was played in the remaining condition. Finally, the participants filled the interaction questionnaire as well as a general questionnaire.

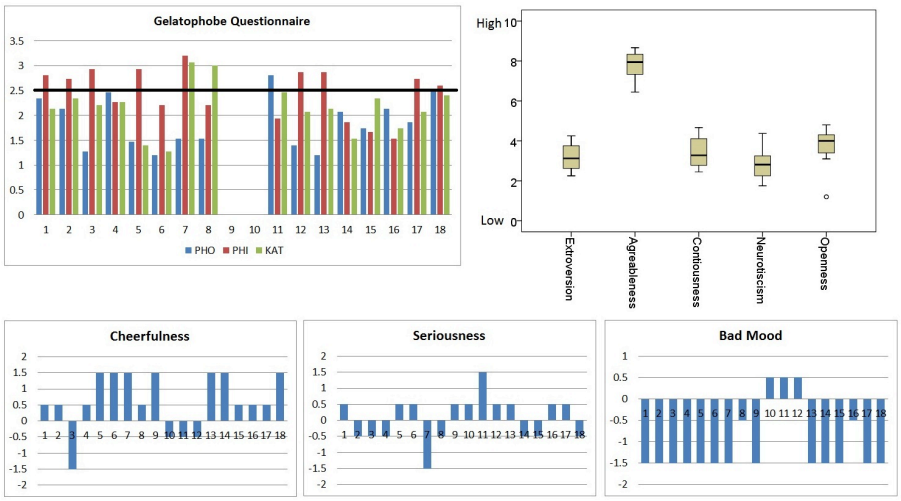


Fig. 12. (Top) Personality trait scores and (Bottom) current mood scores. X-axes indicate the participant number. Participants 9 and 10 did not fill in the personality traits questionnaires (Top). All participants filled in the current mood questionnaire (Bottom).

Results

Figure 12 (top) shows the participants' personality traits in terms of gelotophobia and of extroversion, agreeableness, consciousness, neuroticism, and openness. Only 2 participants scored above the threshold for gelotophobia ($PHO > 2.5$). The general mood (Figure 12 - bottom) was also measured as it could have an effect on the perception of the avatar during the experiment. The figure shows that the participants were overall in a good mood with only three participants scoring high in bad mood.

Figure 13 shows the level of familiarity with and the general likeability of avatars reported by our participants before starting the experiments. We can see from the boxplot for Q4 that our participants present a quite varied level of familiarity with avatars with most of them scoring in the lower part of the scale. The scores for the other questions are also quite low. Only Q2 ("Generally, I enjoy interacting with avatars") and Q5 ("Generally I find interacting with avatars aversive", score inverted for reporting) obtained quite high scores. This shows that, in general, our participants did not dislike interacting with avatars but they had a low confidence in the capabilities that avatars can exhibit when interacting with people.

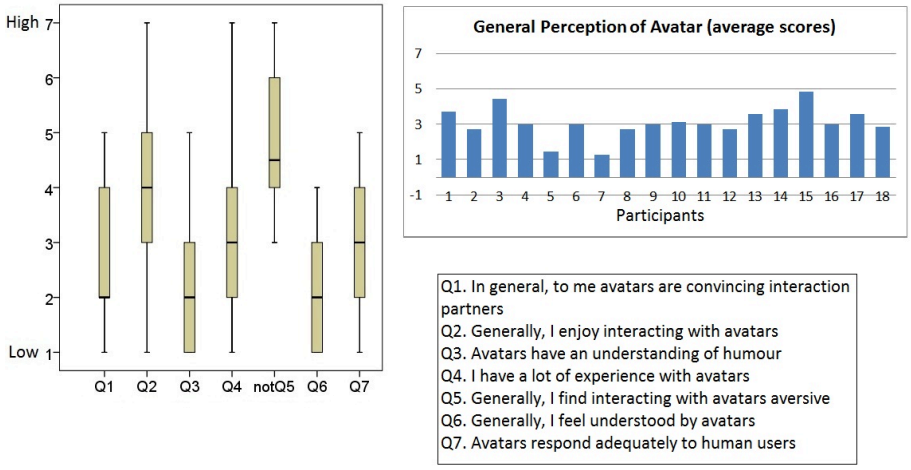


Fig. 13. General familiarity with and perception of likeability and competence of avatars. (Left) scores organized by question; (bottom-right) Q1-Q7 questions. "notQ5" indicates that the response has been inverted for reporting.; (top-right) average scores over the 7 questions for each participant.

In order to identify possible effect of laughter expression on the perception of the avatar, the questions from the in-session questionnaires were grouped into three factors: competence, likeability, naturalness. Naturalness was also separately explored with respect to: naturalness of the non-verbal expressions (excluding laughter-related questions) and of laughter expressions. The grouping of the questions was as follow:

- Competence: Q11, Q13, Q14, Q15, Q17, Q21, Q39
- Likeability: Q12, notQ16, Q18, notQ19, Q20, Q23, Q26, Q27, Q32, Q34, Q35, Q36
- Naturalness: Q22, Q25, Q31, Q37, Q38, Q40, Q41, Q42, Q47, NV, LN (excluding Q24, Q28)
- Non-verbal expressions (NV): Q29, Q30, Q40, Q41, Q42
- Laughter naturalness (LN): Q24, Q28, notQ43, Q44, Q45, Q46

Q24 and Q28 were excluded from the Naturalness factor since many participants did not answer these two questions for the no-laughter condition. These questions were however included in the laughter naturalness factor and a baseline value of 3.5 (middle of the scale) was used when the participant's score was missing.

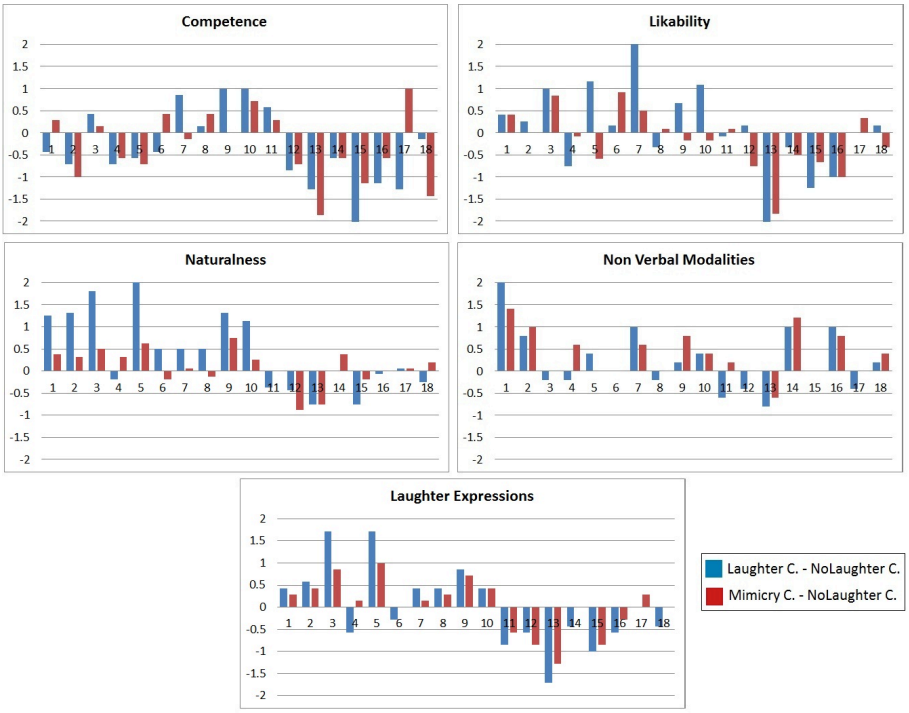


Fig. 14. Comparison of in-session scores between conditions. X-axes indicate participant number; Y-axes indicate the differences between scores obtained in either the laughter condition or the mimicy conditions with respect to the control condition.

The list of questions can be seen in [18]. For each of these factors the scores were normalized. The differences between the laughter condition scores and the no laughter condition scores are shown in Figure 14. The data show high variability between participants' scores. However, some trends can be identified. In particular, the avatar was perceived as a more competent game player in the control conditions than in any of the two conditions with laughter expressions. In the case of likeability, there is a clear split in the participants' reaction to the avatar with many participants reporting greatly increased or decreased liking of the avatar in the laughter conditions compared to the control conditions. A more positive effect is observed in term of naturalness of the avatar.

A repeated-measures test was run to investigate if there were any significant difference between the three conditions. Mauchly's test indicated that the assumption for sphericity was violated for naturalness ($\chi^2(2) = 13.452, p < .01$), non-verbal expression naturalness ($\chi^2(2) = 19.151, p < .01$) and laughter naturalness ($\chi^2(2) = 9.653, p < .001$). Therefore a Greenhouse-Geiesser correction was applied for these three factors. No significant effects were found for the perception of competence, likeability and laughter naturalness. However, significant

effects were found for overall naturalness ($F(1.178, 21.675) = 3.978, p = .05, \mu^2 = .190$) and of non-verbal expression ($F(1.376, 23.4) = 4.278, p = .039, \mu^2 = .201$).

Post hoc comparisons for overall naturalness show that the laughter condition received higher scores than the other two conditions but these differences only approached significance (vs. no-laughter: $p = .15$; vs. mimicry: $p = 1.24$). Post hoc comparisons for non-verbal behaviour show a significant difference ($p = 0.019$) between the no-laughter and mimicry conditions. Figure 15 shows the scores for each of the five questions forming the non-verbal expression factor. We can see that slightly higher scores were obtained for the laughter and mimicry condition with respect to the no-laughter condition. We can also observe higher scores for Q30 for the mimicry condition than for the laughter condition. It is possible that the greater amount of body behaviour (observed in the mimicry condition) may have resulted in the avatar being perceived as more alive. It is also possible that the fact that, in the mimicry condition, the body behaviour was mimicking the body movement of the participants may have captured more their attention. However, only five participants reported feeling that the avatar was mimicking them and only 2 participants correctly indicated in which section the avatar was mimicking and which of the participants was mimicked. In addition, only one person reported that the avatar was mimicking their body movement.

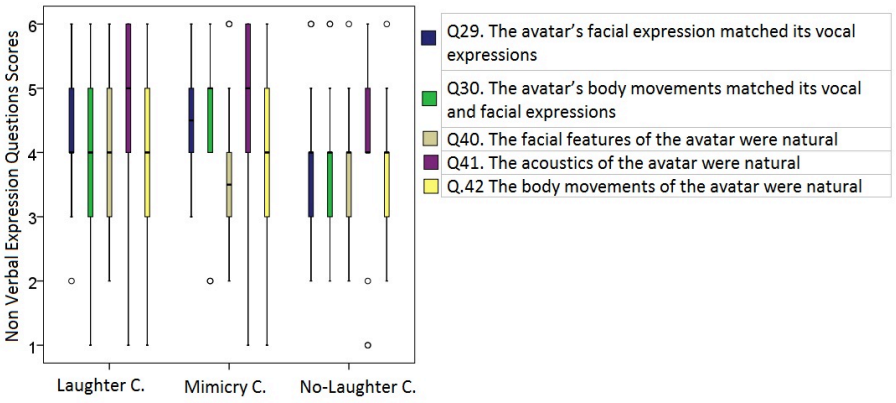


Fig. 15. Boxplots of scores of the questions forming the non-verbal expression factor

The results of this first evaluation showed that laughter added some level of naturalness to the avatar; however, the evaluation also highlighted important technical and experimental design issues that will be addressed before running the full evaluation. In particular, because of the open audio production the avatar detected itself laughing and was unable to distinguish this from participant laughter, it then used this as a cue for generating ever-increasing laughter resulting at times in perceived random or hysterical laughter.

Some technical issues with the synthesis were also identified that need to be addressed to increase naturalness and facilitate communication (e.g., speech synthesis software). Comments from the participants were also very useful and highlighted different problems and solutions to address them. The scenario needs to be slightly redesigned to make sure that the position of the avatar in the triad is more central and participants do not exclude it from the game). Some Wizard of Oz techniques will be used to specifically evaluate individual modules of the laughter machine architecture (e.g., the mimicry module) to avoid the effect being masked by other technical issues (e.g., imperfect recognition of laughter, or lack of natural language understanding).

9 Conclusions

The *Laugh when you're winning* project was designed in the framework of the EU Project ILHAIRE, and its development took place during the eNTERFACE 2013 Workshop, where several partners joined to collaborate for the project setup. Further, the participation in the eNTERFACE Workshop allowed researchers to recruit participants for the testing phase. Tests showed that virtual characters laughter capabilities helped to improve the interaction with human participants. Further, some participants reported that they perceived whether the virtual character was mimicking their behavior.
Several critical points emerged from the project set up and testing and will be addressed in the future:

- the fused detection module is more robust than the one developed in eNTERFACE'12, but on the other hand its reaction time is slightly longer (1-2s) which can cause disturbing delays in the agent's actions; in particular, the agent should not speak simultaneously to the participants but would do so due to the introduced delay; this will be adressed in the future by consulting a low-delay voice activity detection feature when to decide if the agent can speak;
- the cheap microphones used were insufficient for the desired open scenario (agent audio rendered by loudspeakers), which created long laughter loops by the agent; high-quality directional microphones must be used in the future, or the audio of the agent should be rended through headphones;
- the open-source speech synthesis system used with the Greta agent was not intelligible enough, which, in addition to bad timing of some reactions, lead some users to neglect the agent; a professional speech synthesis system will be used in the future to limit this problem;
- more voice/face/body features must be detected or improved; in parallel, the detected features should be synthesised by the virtual character;
- analysis of mimicry during human-human interaction is in progress on the data corpora recorded in the framework of the EU Project ILHAIRE; results will contribute to improved human-virtual character interaction.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement $n°270780$.

References

1. Bachorowski, J., M.J., S., M.J., O.: Automatic discrimination between laughter and speech. Journal of the Acoustical Society of America 110, 1581–97 (2001)
2. Becker-Asano, C., Ishiguro, H.: Laughter in social robotics - no laughing matter. In: International workshop on social intelligence design (SID2009). pp. 287–300 (2009)
3. Becker-Asano, C., Kanda, T., Ishi, C., Ishiguro, H.: How about laughter? perceived naturalness of two laughing humanoid robots. In: Affective Computing and Intelligent Interaction and Workshops, 3rd International Conference on Affective Computing & Intelligent Interaction. pp. 1–6 (2009)
4. Bernhardt, D., Robinson, P.: Detecting affect from non-stylised body motions. In: Affective Computing and Intelligent Interaction, pp. 59–70. Springer (2007)
5. Bourgeois, P., Hess, U.: The impact of social context on mimicry. Biological psychology 77(3), 343–352 (2008)
6. Brand, M.: Voice puppetry. In: Proceedings of conference on Computer graphics and interactive techniques. pp. 21–28 (1999)
7. Bregler, C., Covell, M., Slaney, M.: Video rewrite: driving visual speech with audio. In: Proceedings of the 24th annual conference on Computer graphics and interactive techniques. pp. 353–360. SIGGRAPH '97, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA (1997), http://dx.doi.org/10.1145/258734.258880
8. Cai, R., Lu, L., Zhang, H., Cai, L.: Highlight sound effects detection in audio stream. In: Proceedings of the 2003 IEEE International Conference on Multimedia and Expo (ICME). pp. 37–40. Baltimore, USA (2003)
9. Castellano, G., Villalba, S.D., Camurri, A.: Recognising human emotions from body movement and gesture dynamics. In: Affective computing and intelligent interaction, pp. 71–82. Springer (2007)
10. Cohen, M.M., Massaro, D.W.: Modeling coarticulation in synthetic visual speech. In: Models and Techniques in Computer Animation. pp. 139–156. Springer-Verlag (1993)
11. Cosker, D., Edge, J.: Laughing, crying, sneezing and yawning: Automatic voice driven animation of non-speech articulations. In: Proc. of Computer Animation and Social Agents (CASA09). pp. 21–24. Citeseer (2009)
12. Deng, Z., Lewis, J., Neumann, U.: Synthesizing speech animation by learning compact speech co-articulation models. In: Computer Graphics International 2005. pp. 19–25 (2005)
13. DiLorenzo, P.C., Zordan, V.B., Sanders, B.L.: Laughing out loud: control for modeling anatomically inspired laughter using audio. ACM Transactions on Graphics (TOG) 27(5), 125 (2008)
14. Ezzat, T., Geiger, G., Poggio, T.: Trainable videorealistic speech animation. In: Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on. pp. 57–64 (2004)

15. Fukushima, S., Hashimoto, Y., Nozawa, T., Kajimoto, H.: Laugh enhancer using laugh track synchronized with the user's laugh motion. In: CHI '10 Extended Abstracts on Human Factors in Computing Systems. pp. 3613–3618. CHI EA '10, ACM, New York, NY, USA (2010)
16. Gilroy, S.W., Cavazza, M., Niranen, M., Andre, E., Vogt, T., Urbain, J., Benayoun, M., Seichter, H., Billinghurst, M.: Pad-based multimodal affective fusion. In: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on (2009)
17. Gosling, S., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. Journal of Research in Personality 37(6), 504–528 (2003)
18. Hofmann, J., Platt, T., Urbain, J., Niewiadomski, R., Ruch, W.: Laughing avatar interaction evaluation form. Unpublished research instrument (2012)
19. Kennedy, L., Ellis, D.: Laughter detection in meetings. In: NIST ICASSP 2004 Meeting Recognition Workshop. pp. 118–121. Montreal (May 2004)
20. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: a survey (2012)
21. Kleinsmith, A., Bianchi-Berthouze, N., Steed, A.: Automatic recognition of non-acted affective postures. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 41(4), 1027–1038 (2011)
22. Knox, M.T., Mirghafori, N.: Automatic laughter detection using neural networks. In: Proceedings of Interspeech 2007. pp. 2973–2976. Antwerp, Belgium (August 2007)
23. Kshirsagar, S., Magnenat-Thalmann, N.: Visyllable based speech animation. Comput. Graph. Forum 22(3), 632–640 (2003)
24. Lasarcyk, E., Trouvain, J.: Imitating conversational laughter with an articulatory speech synthesis. In: Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter. pp. 43–48 (August 2007)
25. Leite, I., Castellano, G., Pereira, A., Martinho, C., Paiva, A.: Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction. pp. 367–374. ACM (2012)
26. Liu, W., Yin, B., Jia, X., Kong, D.: Audio to visual signal mappings with hmm. In: In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 04 (2004)
27. Mancini, M., Glowinski, D., Massari, A.: Realtime expressive movement detection using the eyesweb xmi platform. In: Camurri, A., Costa, C. (eds.) INTETAIN. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 78, pp. 221–222. Springer (2011)
28. Mancini, M., Hofmann, J., Platt, T., Volpe, G., Varni, G., Glowinski, D., Ruch, W., Camurri, A.: Towards automated full body detection of laughter driven by human expert annotation. In: Proceedings of the Fifth Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction, Affective Interaction in Natural Environments (AFFINE) Workshop. pp. 757–762. Geneva, Switzerland (2013)
29. Mancini, M., Varni, G., Glowinski, D., Volpe, G.: Computing and evaluating the body laughter index. Human Behavior Understanding pp. 90–98 (2012)
30. Meng, H., Kleinsmith, A., Bianchi-Berthouze, N.: Multi-score learning for affect recognition: the case of body postures. In: Affective Computing and Intelligent Interaction, pp. 225–234. Springer (2011)

31. Niewiadomski, R., Mancini, M., Baur, T., Varni, G., Griffin, H., Aung, M.: Mmli: Multimodal multiperson corpus of laughter in interaction. In: Fourth Int. Workshop on Human Behavior Understanding, in conjunction with ACM Multimedia'2013 (2013)
32. Niewiadomski, R., Pelachaud, C.: Towards multimodal expression of laughter. In: Proceedings of the 12th international conference on Intelligent Virtual Agents. pp. 231–244. Springer-Verlag Berlin, Heidelberg (2012)
33. Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., Cakmak, H., Pammi, S., Baur, T., Dupont, S., Geist, M., Lingenfelser, F., McKeown, G., Pietquin, O., Ruch, W.: Laugh-aware virtual agent and its impact on user amusement. In: Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems. pp. 619–626. AAMAS '13, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2013)
34. Oura, K.: Hmm-based speech synthesis system (hts) [computer program webpage]. http://hts.sp.nitech.ac.jp/ (consulted on June 22, 2011)
35. Owens, M.D.: It's all in the game: Gamification, games, and gambling. Gaming Law Review and Economics 16 (2012)
36. Petridis, S., Pantic, M.: Audiovisual discrimination between laughter and speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5117–5120. Las Vegas, Nevada (2008)
37. Petridis, S., Pantic, M.: Is this joke really funny? Judging the mirth by audiovisual laughter analysis. In: Proceedings of the IEEE International Conference on Multimedia and Expo. pp. 1444–1447. New York, USA (June 2009)
38. Poe, E.A.: Maelzel's chess-player. In: Southern Literary Messenger, vol. 2, pp. 318–326 (1836)
39. Qu, B., Pammi, S., Niewiadomski, R., Chollet, G.: Estimation of FAPs and intensities of AUs based on real-time face tracking. In: Pucher, M., Cosker, D., Hofer, G., Berger, M., Smith, W. (eds.) The 3rd International Symposium on Facial Analysis and Animation. ACM (2012)
40. Ruch, W., Ekman, P.: The expressive pattern of laughter. In: Kaszniak, A. (ed.) Emotion, qualia and consciousness, pp. 426–443. World Scientific Publishers, Tokyo (2001)
41. Ruch, W., Proyer, R.: Extending the study of gelotophobia: On gelotophiles and katagelasticists. Humor-International Journal of Humor Research 22(1/2), 183–212 (2009)
42. Ruf, T., Ernst, A., Küblbeck, C.: Face detection with the sophisticated high-speed object recognition engine (shore). In: Heuberger, A., Elst, G., Hanke, R. (eds.) Microelectronic Systems, pp. 243–252. Springer Berlin Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-23071-4_23
43. Scherer, S., Glodek, M., Schwenker, F., Campbell, N., Palm, G.: Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. ACM Trans. Interact. Intell. Syst. 2(1), 4:1–4:31 (Mar 2012)
44. Sundaram, S., Narayanan, S.: Automatic acoustic synthesis of human-like laughter. In: Journal of the Acoustical Society of America. vol. 121, pp. 527–535 (January 2007)
45. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for hmm-based speech synthesis. In: ICASSP. pp. 1315–1318 (2000)

46. Tokuda, K., Zen, H., Black, A.: An HMM-based speech synthesis system applied to English. In: Proceedings of the 2002 IEEE Speech Synthesis Workshop. pp. 227–230. Santa Monica, California (September 2002)
47. Truong, K.P., van Leeuwen, D.A.: Automatic discrimination between laughter and speech. Speech Communication 49, 144–158 (2007)
48. Urbain, J., Çakmak, H., Dutoit, T.: Arousal-driven synthesis of laughter. submitted to the IEEE Journal of Selected Topics in Signal Processing, Special Issue on Statistical Parametric Speech Synthesis (2014)
49. Urbain, J., Cakmak, H., Dutoit, T.: Development of hmm-based acoustic laughter synthesis. In: Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalisations in Speech. pp. 26–27. Dublin, Ireland (2012)
50. Urbain, J., Cakmak, H., Dutoit, T.: Evaluation of hmm-based laughter synthesis. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, Canada (May 2013)
51. Urbain, J., Dutoit, T.: A phonetic analysis of natural laughter, for use in automatic laughter processing systems. In: Proceedings of the fourth bi-annual International Conference of the HUMAINE Association on Affective Computing and Intelligent Interaction (ACII2011). pp. 397–406. Memphis, Tennesse (October 2011)
52. Urbain, J., Niewiadomski, R., Bevacqua, E., Dutoit, T., Moinet, A., Pelachaud, C., Picart, B., Tilmanne, J., Wagner, J.: 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 4(1), 47–58 (2010), special Issue: eNTERFACE'09
53. Urbain, J., Niewiadomski, R., Mancini, M., Griffin, H., Huseyin Cakmak, H., Ach, L., Volpe, G.: Multimodal analysis of laughter for an interactive system. In: Proceedings of the INTETAIN 2013. Mons, Belgium (2013)
54. Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., André, E.: The social signal interpretation (ssi) framework - multimodal signal processing and recognition in real-time. In: Proceedings of the 21st ACM International Conference on Multimedia, 21-25 October 2013, Barcelona, Spain. (2013)