J Multimodal User Interfaces (2010) 4: 47–58
DOI 10.1007/s12193-010-0053-1

ORIGINAL PAPER

AVLaughterCycle

Enabling a virtual agent to join in laughing with a conversational partner using a similarity-driven audiovisual laughter animation

Jérôme Urbain · Radoslaw Niewiadomski · Elisabetta Bevacqua · Thierry Dutoit · Alexis Moinet · Catherine Pelachaud · Benjamin Picart · Joëlle Tilmanne · Johannes Wagner

Received: 26 February 2010 / Accepted: 11 September 2010 / Published online: 28 September 2010

Abstract The AVLaughterCycle project aims at developing an audiovisual laughing machine, able to detect and respond to user's laughs. Laughter is an important cue to reinforce the engagement in human-computer interactions. As a first step toward this goal, we have implemented a system capable of recording the laugh of a user and responding to it with a similar laugh. The output laugh is automatically selected from an audiovisual laughter database by analyzing acoustic similarities with the input laugh. It is displayed by an Embodied Conversational Agent, animated using the audio-synchronized facial movements of the subject who originally uttered the laugh. The application is fully implemented, works in real time and a large audiovisual laughter database has been recorded as part of the project.

This paper presents AVLaughterCycle, its underlying components, the freely available laughter database and the application architecture. The paper also includes evaluations of several core components of the application. Objective tests show that the similarity search engine, though simple, significantly outperforms chance for grouping laughs by speaker or type. This result can be considered as a first measurement for computing acoustic similarities between laughs. A subjective evaluation has also been conducted to measure the influence of the visual cues on the users' evaluation of similarity between laughs.

Keywords Laughter · Embodied Conversational Agent · Acoustic similarity · Facial motion tracking

1 Motivation and related work

Laughter is an essential signal in human communication. It conveys information about our affects and helps to cheer up our mood. Moreover, it is contagious, eases social contacts and has the potential to elicit emotions in listeners. Laughter is also known to have healthy effects, and especially as of the best remedies for stress [3]. Many events connecting and entertaining people from all over the world through the universal signal of laughter have been successful, like the World Laughter Day or the Skype Laughter Chain [28].

Due to the growing interest for virtual entities modeling human behaviors, a need to enable these machines to perceive and express emotions has emerged. Laughter is clearly an important cue for understanding affects and discourse events as well as creating affects and providing feedback to the conversational partners. There is a strong interest for integrating laughter in human-computer interaction, for example for educational devices [30]. In consequence, automatic laughter processing has gained in interest during the last decades. Laughter is considered as a raw affect burst, "expected to be barely conventionalized, thus relatively universal, and show strong inter-individual differences" [26]. Indeed, laughter is a highly variable signal and it is hard to describe its acoustic structure. Trouvain [33] summarizes

Portions of this work have been presented in "Proceedings of eNTERFACE'09" [36].

J. Urbain (✉) · T. Dutoit · A. Moinet · B. Picart · J. Tilmanne
Faculté Polytechnique de Mons, TCTS Lab, Université de Mons,
20, Place du Parc, 7000 Mons, Belgium
e-mail: jerome.urbain@umons.ac.be

R. Niewiadomski · E. Bevacqua · C. Pelachaud
CNRS-LTCI UMR 5141, Institut TELECOM–TELECOM
ParisTech, 37/39, rue Dareau, 75014 Paris, France

J. Wagner
Institut für Informatik, Human-Centered Multimedia,
Augsburg University, Universitätsstr. 6a, 86159 Augsburg,
Germany

the different terminologies used in previous laughter studies, as well as various categories to designate laughter types.

On the automatic recognition side, efficient systems to discriminate between laughter and speech have been developed. Truong and van Leeuwen [34] compared several audio feature sets and classifiers for distinguishing segments of speech and laughter. Using Perceptual Linear Prediction Coding and prosodic features and fusing, via a Multi-Layer Perceptron (MLP), the outputs of a Support Vector Machine and a Gaussian Mixture Model classifier, they obtained an Equal Error Rate (EER—the lower the EER, the better the performance) of 3% on the ICSI Meeting Corpus [11]. The EER rises to 11% when classifying unsegmented laughter in raw meeting files [35]. Knox and Mirghafori [12] obtained slightly better results (8%), also combining spectral (MFCCs) and prosodic features. Their classification was based on MLPs fed with a current feature vector as well as contextual features. Petridis and Pantic [22] combined acoustic (spectral and prosodic) and visual features to discriminate between segments of speech, voiced and unvoiced laughter with a 75% accuracy.

Acoustic laughter synthesis is a complex task. Sundaram and Narayanan [31] state that a good model for laughter synthesis should: (1) be able to generate a broad range of laughs, varying in durations or sounds, as people do; (2) produce human-like variations of characteristic parameters, inside a laugh, otherwise it will not be judged as natural; (3) enable to synthesize laughs providing simple information. Modeling the laughter energy envelope with a mass-spring oscillation and synthesizing laughter vowels by Linear Prediction, Sundaram and Narayanan generated computer laughs, but these were judged as non-natural by listeners. Lasarcyk and Trouvain [14] compared laughs synthesized by an articulatory system (a 3D modeling of the vocal tract) and diphone concatenation. The articulatory system gave better results, but they were still evaluated as significantly less natural than human laughs.

The recent technological progress has made the creation of a humanoid interface to computer systems possible. An Embodied Conversational Agent (ECA) is a computer-generated animated character able to carry on natural, human-like communication with users. In the last twenty years several ECA architectures were developed both by the research community (e.g. [5, 13]) and industry (e.g. [4, 9]). There are few works on synthesis of laugh for virtual agents [7] and robots [2]. Nijholt [17] discusses the advantages and difficulties of introducing humor and laughter to embodied agents, while Becker-Asano and Ishigure [2] evaluate the role of the laughter in the perception of social robots.

The aim of the AVLaughterCycle project is to endow a virtual agent with the capability to join its conversational partner's laugh. Given the difficulty of generating laughter, it was decided not to synthesize laughs but to have the virtual agent, Greta [16], display an unmodified audiovisual human laughter. Laughs are automatically selected from a large database, which contains a broad range of laugh sounds and durations.

The functionality of AVLaughterCycle application can be divided in three tasks:

- Building a large audiovisual database of spontaneous human laughs.
- Properly animating the virtual agent's face movements simultaneously with the laughter acoustics, by transposing captured facial motions to the virtual agent's morphology.
- Selecting a laugh that should be played in answer to the user's laugh, using an organization of laugh similarity.

Integrating these three tasks, the AVLaughterCycle application enables the user to laugh and see Greta laughing in response. When a non-silent activity is detected, with the assumption it is laughter, AVLaughterCycle looks for the most similar laugh in the AVLaughterCycle database and instructs the Greta agent to display it immediately. Users can thus experience a "laughter dialog" with Greta.

The AVLaughterCycle application will also serve us to consider the audiovisual aspect of laughs. Frequently in laughter processing, laughter is regarded as an acoustic act only. In this work we argue that laughter is a behavior containing contributions from both acoustic and visual components. A vast overview of the visual cues of laughter can be found in [24]. The laughter expressions are multimodal and are composed of several facial movements (e.g. zygomaticus major, levator labii superior is, depressor anguli oris) as well as body movements (e.g. backward tilt of the head, shaking of the shoulders). Indeed we will show that the non-acoustic behaviors that are synchronized with the audio cues are a significant part of the experience of laughter.

Besides enabling the selection of a laugh in the current application, grouping laughs by similarities can be beneficial for other fields like laugher recognition, laughter database browsing or laughter classification. Computing acoustic similarities between laughs (other than the discrimination between voiced and unvoiced) is an innovative approach, hence the evaluation presented in this paper can be considered as a first measurement of this type, and a baseline to evaluate future algorithms.

The paper is organized as follows. Section 2 is dedicated to the software used during this project: Smart Sensor Integration (SSI) [39] for recording, annotating and analyzing laughs; MediaCycle [27] to compute similarities between laughs; Greta [16] for playing the output laughter. Section 3 presents the audiovisual laughter database, that contains laughs used to animate Greta. Section 4 describes the AVLaughterCycle application process and its methods for analyzing the input laughter, selecting an answering laughter and driving Greta accordingly. Section 5 focuses on the evaluation of the system. Finally, conclusions and future works are presented in Sect. 6.

2 Integrated software

2.1 Smart Sensor Integration

Smart Sensor Integration is software designed to deal with multimodal signal recording and processing. It provides a Graphical User Interface (GUI) to start and stop a recording. The GUI includes a dedicated space to present stimuli, which is useful for database recordings. Afterwards the data can be visualized and annotated. The different modalities (speech, video, etc.) are automatically synchronized.

SSI integrates signal processing libraries and many signal processing algorithms can be interfaced with it. Input signals can be analyzed in real-time or offline.

In this project, SSI was used to manage the database recordings and annotate them (see Sect. 3), as well as for processing the audio input and computing acoustic features in our real-time application (see Sect. 4).

2.2 MediaCycle

MediaCycle is software developed for browsing through large multimedia databases, using similarity. It started by considering acoustic similarities only, in a project called AudioCycle [8], designed to ease the navigation inside musical audio sample databases. The software computes acoustic features—characterizing musical properties of rhythm, melody and timbre—for each file in an audio sample database and then evaluates the similarities between samples through the (weighted) distances between their feature vectors.

AudioCycle has been extended in a project called MediaCycle where image, video and laughter features were added. The system can be queried by laughing; the incoming laughter is placed in the database space and the N most similar laughs are returned.

2.3 The 3D humanoid agent: Greta

Greta [16] (Fig. 1) is a 3D humanoid agent. It is able to communicate with the user using verbal and nonverbal channels like gaze, facial expressions and gestures. It follows the SAIBA framework [38] that defines a modular structure, functionalities and communication protocols for Embodied Conversational Agents (ECAs) and the MPEG-4 [18] standard of animation. Greta is a complex architecture composed of several modules (i.e. Intent Planner, Behavior Planner, Behavior Realizer, Player; see [16] for details) that uses two XML-languages: FML-APML [10] and BML [38]. Recently it has been equipped with four different characters.

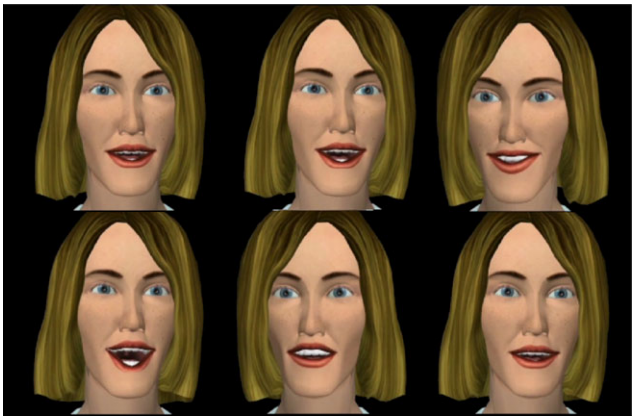

Fig. 1 Greta, the 3D humanoid agent used in AVLaughterCycle, laughing

In the AVLaughterCycle application we are using only some parts of the Greta agent and BML language that specifies its verbal and nonverbal behaviors. Each BML tag corresponds to a behavior the agent has to produce on a given modality: head, torso, face, gaze, gesture, speech. These signals are sent to the Behavior Realizer module that generates the MPEG-4 Face and Body Animation Parameter (FAP-BAP) files. Finally, the animation is played in the FAP-BAP Player. All modules in the Greta architecture are synchronized using a central clock and communicate with each other through the Psyclone messaging system [32]. The system has a low latency time that makes it suitable for interactive applications. Prior to this project, Greta was able to display a variety of nonverbal affective behaviors, but not laughter.

3 Creation of an AV laughter database

The first step of the AVLaughterCycle project was the recording of an audiovisual (AV) database consisting of humans laughing. Only that database information required to understand this paper is presented here. More details about the database (recording protocol, stimuli, annotation, contents) can be found in [37]. The database is freely available on http://tcts.fpms.ac.be/~urbain.

24 subjects (9 females, 15 males) participated in the database. Laughter was elicited by a 10-minute comedy video. Subjects wore a headset microphone for stimulus listening and audio-recording of their reactions (16 kHz, PCM 16 bits). In addition, facial motion tracking was performed. Although automatic markerless techniques for detecting facial actions are emerging and proving efficient in the emotion recognition and behavior science fields [1, 19, 25], it is still extremely difficult to deal with 3D, continuous (compared to a binary decision for each Action Unit) represen-

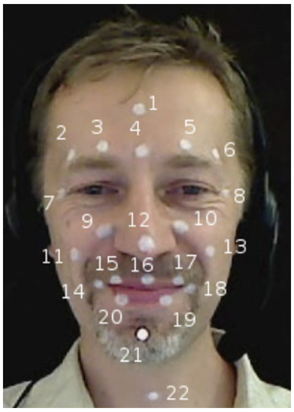

Fig. 2 ZignTrack—22 face markers

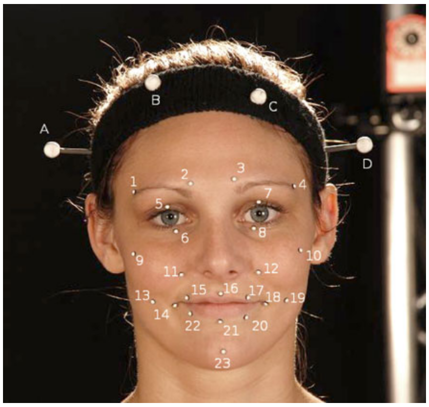

Fig. 3 OptiTrack—23 face (1–23) and 4 head (A–D) markers [15]

tation of spontaneous expressions. In consequence, marker-based techniques were preferred for this project, a choice often made for realistic 3D avatar animation. Two commercial systems have been used for building the database: 8 participants (3 females, 5 males) were recorded with ZignTrack [41], using 22 markers (Fig. 2) tracked by a simple webcam (25FPS); and 16 participants (6 females, 10 males) were recorded with the OptiTrack setup [15], consisting of 27 infrared markers on the face and head (Fig. 3) tracked by 6 100FPS infrared cameras placed in a hemisphere (Fig. 4; the 7th camera, in the center, is not used for motion tracking but for scene recording). Data is immediately recorded in 3D with OptiTrack, while it is extrapolated from 2D using a fixed face template by the ZignTrack software.

From the 24 recordings, 1021 laughs were labeled. The annotation protocol was designed to help refine the laughter description. In addition to a main class (laugh, verbal, breath, trash), a label could be extended to add details about the segment contents. The extended label is used mostly for the following details of the laugh class:

- The laughter temporal structure—following the three segmentation levels presented by Trouvain [33], these "sublabels" indicate whether the *episode* (i.e., the full laughter utterance) contains several *bouts* (i.e. parts separated by inhalations), only one, or only one syllable. These temporal structure sublabels are mutually exclusive.
- The laughter acoustic contents—sublabels refer to the type of sound: vowel, breathy (breathing sound traveling through the open mouth), nasal (breathing sound traveling through the nose), grunt-like, hum-like, "hiccup-like", speech-laughs or barely audible (quasi-silent). The sublabels can be combined to reflect a change in the acoustic content during the laughter, for example, if a laugh starts with grunt-like sounds and is followed by nasal respiration sounds, it receives the label "laugh_grunt_nasal".

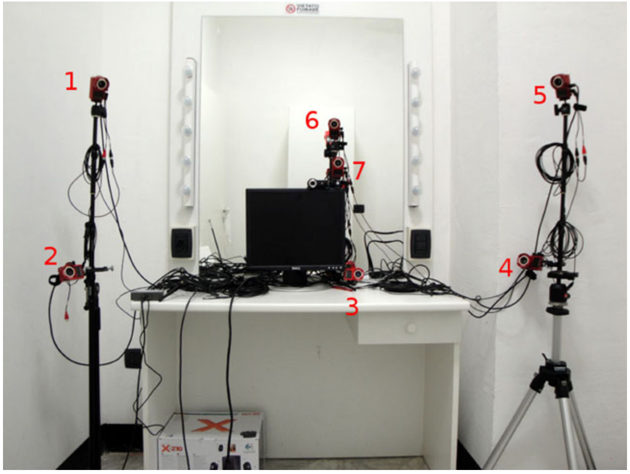

Fig. 4 OptiTrack—7 infrared cameras

Table 1 Occurrences of the main classes

Main class	Occurrences
Laugh	1021
Trash	207
Verbal	64
Breath	31

Table 1 gives the number of occurrences of each of the main classes over the whole database. The number of occurrences of the laughter subclasses is presented in Table 2. There are more acoustic sublabels (1356) than number of laughs (1021), because one laugh can receive several acoustic sublabels.

4 Corpus based AudioVisual Laughter synthesis

The communication between the different modules of the AVLaughterCycle application is illustrated in Fig. 5. Users can query the system in two ways: by sending a full audio laughter file (offline mode) or in real-time (online mode), using SSI for recording and real-time processing. In the latter case, SSI segments the audio input by thresholding the

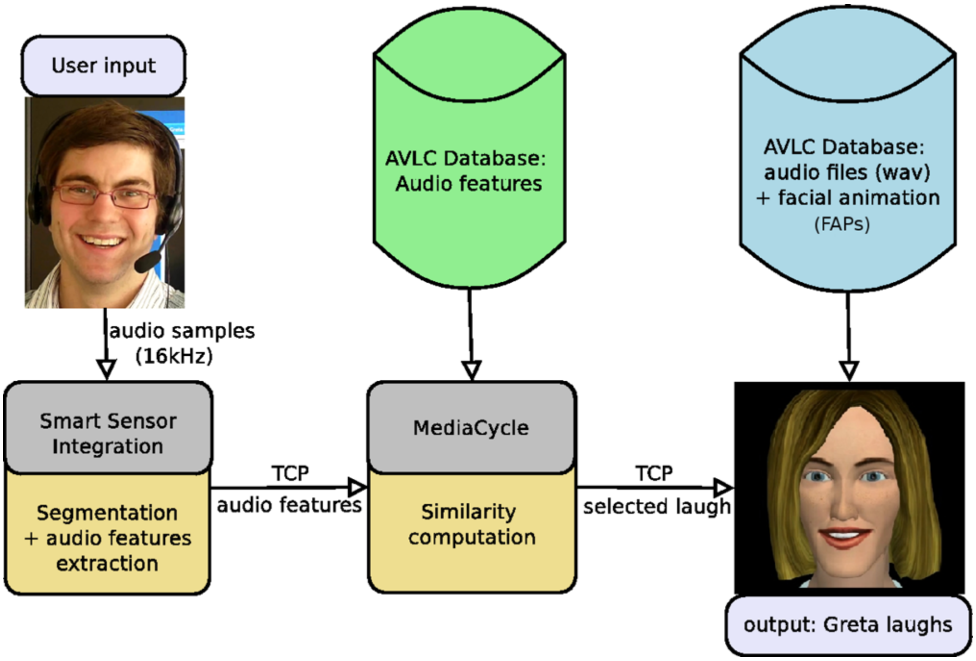

Fig. 5 Flow chart of the AVLaughterCycle application

Table 2 Occurrences of the laughter subclasses

Category	Laughter subclass	Occurrences
Structure	Monosyllabic	179
	One *bout*	677
	Several *bouts*	165
Acoustic	Vowel	446
	Nasal	277
	Breath	237
	Hum	169
	Hiccup	95
	Grunt	18
	Co-occurring speech and laugh	20
	Silent	94

signal to noise ratio (SNR). There is no laughter detection for the moment: input is assumed to be laughter and every segment satisfying the SNR condition is further processed. In both online and offline modes, when the audio laugh segment is available, SSI computes its features (see Sect. 4.1) and sends them to MediaCycle. MediaCycle compares these features with the database samples and returns the most similar laugh. This laugh is sent to Greta, which plays the audio sound synchronously with the corresponding facial animation (see Sects. 4.2 and 4.3). Greta answers immediately after the end of the user's laugh was detected.

4.1 Laughter audio similarity analysis

The labeled laugh segments are all processed by the MediaCycle tool to compute their similarities and cluster them. MediaCycle evaluates the similarities by measuring distances between feature vectors. Features have been based on Peeters's set [21] and implemented in C++. The features can be extracted directly in SSI, in which the MediaCycle audio feature extraction library has been integrated, and then sent to MediaCycle. In the current version of AVLaughterCycle, the following spectral features are used:

- 13 Mel-Frequency Cepstral Coefficients (MFCCs), and their first and second derivatives.
- Spectral flatness and spectral crest values, each divided in 4 analysis frequency bands (250 Hz to 500 Hz, 500 Hz to 1000 Hz, 1000 Hz to 2000 Hz and 2000 Hz to 4000 Hz).
- Spectral centroid, spread, skewness and kurtosis.
- Loudness, sharpness and spread, computed on the Bark frequency scale.
- Spectral slope, decrease, roll-off and variation.

In addition, 2 temporal features are included: the energy and the zero-crossing rate. In total, 60 features are extracted for each frame of 340 samples (Fs: 16 kHz), with 75% overlap. The similarity estimation requires comparing audio segments of different lengths, so also comparing different numbers of frames. To obtain a constant feature vector size, it was decided to store only the mean and standard deviation of each feature over the whole segment. More complex models could be investigated but this simple transform provides promising results and establishes a baseline, useful to measure future improvements. This simplification had been successfully used in other similarity computation [8] or laughter classification [22] contexts and was assumed applicable to laughter timbre characterization. Normalized Euclidean distance between feature vectors is used to compute the similarity between laughter episodes.

When AVLaughterCycle is queried, the input laugh is analyzed and its audio feature vector is computed. This vector is used to select a corresponding laugh inside the laughter database. In this project, it was decided to return the closest (i.e. most similar according to our feature set) laugh from the input laugh. Doing so, the system can be employed to search inside the database for a specific kind of laughter.

4.2 Visual replay

The transfer from motion capture data to MPEG-4 FAPs is divided in two steps. We follow the common procedure that is used in Computer Graphics [20, 23]. First, we extract facial movements from motion capture data, then we perform retargeting to our MPEG-4 model.

The data from the motion capture software contains, for each frame, the position of each marker. These values express the absolute distance of the markers in 3D space to the predefined central point. The positions of the markers on the face are modulated by head rotations and body movements. The information about the head and body movements is available both in ZignTrack (in BVH format) and OptiTrack (in BVH and C3D formats) and was discarded to keep only the facial movements. Noise caused by the technical flaw of the capturing hardware was removed using frequency filters. All the values were then converted to express only the relative movements i.e. the movements in relation to the neutral expression. For each point we subtracted its value from the first frame of the video showing the neutral expression.

This type of data was used to animate Greta based on the MPEG-4 standard. In this standard, the face model is animated by 66 parameters called FAPs. Each parameter deforms one region of the face in one direction (i.e. horizontal or vertical). The automatic parametrical retargeting of the spontaneous facial behavior to a geometric model is still an open issue and several approaches have been proposed (see [6, 29, 40]). Spontaneous laughter behaviors that involve many rapid and short movements add difficulty to the automatic retargeting procedure.

Our database has been elaborated using two motion capture systems: one in 2D, the other in 3D. They have different numbers of markers. Moreover, these systems do not have a high number of markers which makes it difficult to capture all the subtle facial expressions. Advanced commercial systems often use more than 300 markers on the face. Our database also encompasses a large variety of subjects, each with their own facial shape. With all these configurations in mind, we opted for manual retargeting. Thus for each of the motion capture systems, we built an interpolation function between the facial markers and the MPEG-4 facial parameters. For some of the markers, this interpolation is straightforward but some FAPs do not have any corresponding markers (e.g. for the inner lip parameters). In such a case, we defined some extrapolation mappings.

Let us consider some examples of the mappings we used. In the data generated by ZignTrack, no marker corresponds to FAP 5 (raise_b_midlip)—the value of this point is calculated by the arithmetic mean of the two other markers located on the lower lip (19 and 20 on Fig. 2). Similarly, FAPs 37 and 38 (squeeze_l_eyebrow and squeeze_r_eyebrow) do not have correspondence with ZignTrack markers. To estimate the value of FAPs 37 and 38 we use the marker placed at the middle of the appropriate eyebrow (3 and 5 on Fig. 2). If the y coordinate of this marker is positive, the value FAP 37 or 38 increases proportionally to y, otherwise it is 0.

To avoid unnatural facial expressions we also added some constraints on the FAPs values. For example, FAPs 55 and 56 (lower_t_lip_lm_o and lower_t_lip_rm_o) cannot have higher value than FAP 51 (lower_t_midlelip) that is placed between them. To overcome some limitations of MPEG-4 due to a small number of parameters in some facial areas (e.g. cheek), we have defined one-to-many mappings, for example, AU6 (orbicularis oculi activity), which often occurs in spontaneous laughter [24], is difficult to simulate with FAPs. We do so with the partial closure of the lower eyelids (FAPs 21 and 22) and the horizontal displacement of the cheeks (FAPs 39 and 40).

4.3 Greta laughs

Our laughter database contains the precise, frame-by-frame descriptions of partial animations (i.e. only the face) in FAP format. However, animation generation in Greta's engine is by default realized within the procedural approach: Greta's verbal and nonverbal behaviors are defined in BML language (see Sect. 2.3), and single nonverbal behaviors are defined using high level symbolic representation.

In this project, the default BML syntax has been extended to allow mixing (high level) BML commands with (low level) FAPs description. Greta's animation engine was modified to be able to generate smooth animation for such content. Consequently Greta may display a laughter animation using the data from the laughter database, which is accompanied by an audio file and other nonverbal signals that might be specified in BML language (like gestures). When motion capture driven facial animation and non facial procedural animation (i.e. gestures, gaze, torso and head movements) overlap in time, both are displayed simultaneously, without conflict. When two conflicting facial animations are to be displayed at the same time, the motion capture has a higher priority than the procedural facial animation. In such a case, to ensure the final animation remains smooth, the engine interpolates between the first and the last frame of the motion capture driven animation and the adjacent key frames of the procedural one.

Greta was integrated in the AVLaughterCycle using Psyclone software and BML commands. The AVLaughterCycle database (see Sect. 3) contains, for each audio sample, the original motion capture data and the resulting FAPs values from manual conversion. MediaCycle uses Psyclone to send the BML command containing the reference to an audio laugh file and the corresponding visual data in MPEG-4 format.

5 Evaluation

Evaluating the whole AVLaughterCycle application is quite a difficult task and it requires measuring how well it responds to incoming laughs. The only way to evaluate the application is through perceptive tests. A global analysis of the application performance would merge the output selection and the animation, so it would be hard to know what is effective and what is not. It was decided to evaluate several core blocs of the application separately and use objective measures when possible. The face motion tracking systems are briefly compared in Sect. 5.1. Then, objective measures assessing the efficiency of MediaCycle are presented in Sect. 5.2. Finally, Sect. 5.3 describes a subjective experiment to evaluate the influence of the animation on users' ratings of similarity.

5.1 Comparison of the two face motion tracking systems

The two face motion capture systems we used are notably different. Although OptiTrack is low-priced compared to professional systems used in 3D films production, it costs 40 times more than ZignTrack. It is thus not surprising that the comparison favors OptiTrack.

ZignTrack works quite well if markers stay visible during the whole recording and head movements are slow. The tracking fails otherwise, which requires heavy manual corrections. Unfortunately, this happened often in our laughter recordings. In addition, because the 3D extrapolation from 2D uses a fixed face template, distortions occur when there are head rotations.

The OptiTrack software performed better. Even when some markers are lost during the tracking (due to extreme head rotations), they are nearly always recovered after a short time, due to the 6 points of view and infrared (versus visible spectrum) acquisition performance. About 25 minutes of manual post-processing were required to check and adjust the real 3D position of 27 markers for each 10-minute, 100FPS recording.

5.2 MediaCycle

To objectively evaluate the MediaCycle similarity estimation, two experiments were conducted. First, since the current similarity computation is based on spectral features characterizing the timbre of a laugh, the capability of MediaCycle to group laughs from the same speaker was estimated. Second, we measured how often the most similar laughs chosen by MediaCycle contain the same label as the input laughter. For these tests, some laughs were discarded from the database: laughs involving speech (20); 19 laughs from *Subject*1 for which we do not have facial tracking; the laughs from *Subject*24, who only uttered 4 short laughs, which is not enough to perform reliable tests. In total, these experiments involved 978 laughs.

5.2.1 Laughter-based speaker recognition

Each laugh in the database was given as input to MediaCycle, which returned the N closest neighbors. If at least one of the N outputs had been uttered by the input speaker, the MediaCycle search was considered successful. Figure 6 gives the individual success rates for $N = 1, 3, 5$ and 10. The gray bar represents the likelihood of a successful search if randomly selecting laughs instead of using MediaCycle. The random success score for speaker i and N random picks equals

$$R_i^N = 1 - \prod_{k=1}^{N} \frac{N_{tot} - N_i - k + 1}{N_{tot} - k} \qquad (1)$$

where N_i is the number of laughs from speaker i out of the N_{tot} laughs in the database.

For each value of N, MediaCycle performs significantly better than chance, at a 95% confidence level (all p-values are largely lower than 0.05, using one-sided paired t-tests). However, for some individuals (subjects 10, 19 and, to a smaller extent, 8), MediaCycle does not outperform chance. This is probably due to the fact that these subjects mainly uttered nasal or breathy laughs, for which it is very hard to discriminate between subjects (there is no perceived timbre). On the other hand, Subject 11, who gets a (nearly) perfect success rate, produced a large majority of voiced ("vowel") laughs.

To complement this information and illustrate the interest of using MediaCycle to organize a laughter database according to the speaker, we have computed, for each laugh, the average number of utterances one needs to pick to find one laugh from the same speaker. Again, MediaCycle (utterances ordered by distance to the input laugh in the fea-

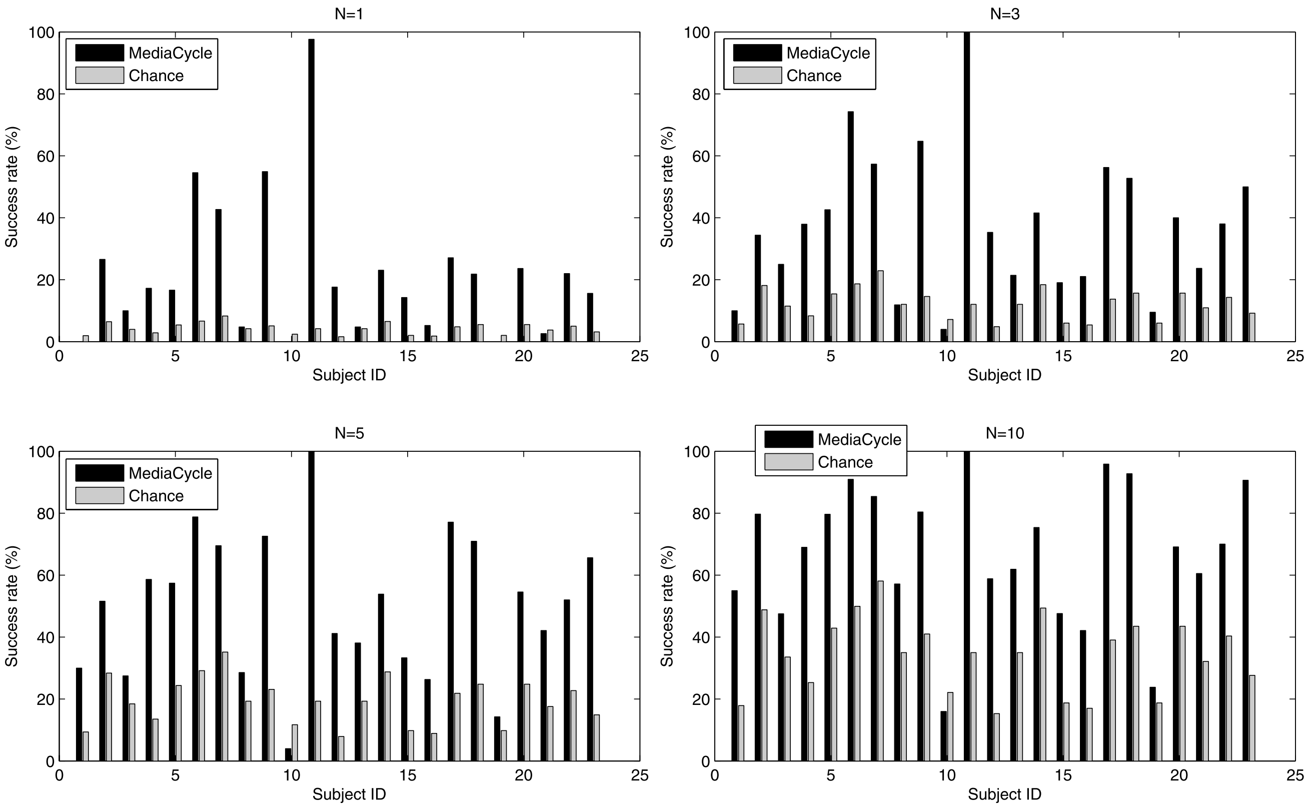

Fig. 6 Success rates achieved by MediaCycle (*black*) against chance (*gray*) for laugher retrieval, using N picks

ture space) was compared against chance. The mean chance score for speaker i equals

$$C_i = \sum_{u=1}^{N_0+1} u \cdot \frac{N_1 - 1}{N_{tot} - u} \prod_{t=1}^{u-1} \frac{N_0 - t + 1}{N_{tot} - t} \tag{2}$$

where N_i is the number of laughs from speaker i out of the N_{tot} laughs in the database and $N_0 = N_{tot} - N_i$ is the number of laughs from other speakers. The results are shown on Fig. 7, with the standard deviation intervals for MediaCycle. MediaCycle is undoubtedly better than random search,[1] though for 5 Subjects (3, 7, 14, 20, 21), the *mean* + *std* value goes above (i.e. is worse than) the chance performance. The one-sided paired t-test gives a p-value of 7.1×10^{-8}. For unknown reasons, the interface was not able to efficiently improve the search for *Subject*3, who uttered 40 laughs spread over the laughs types.

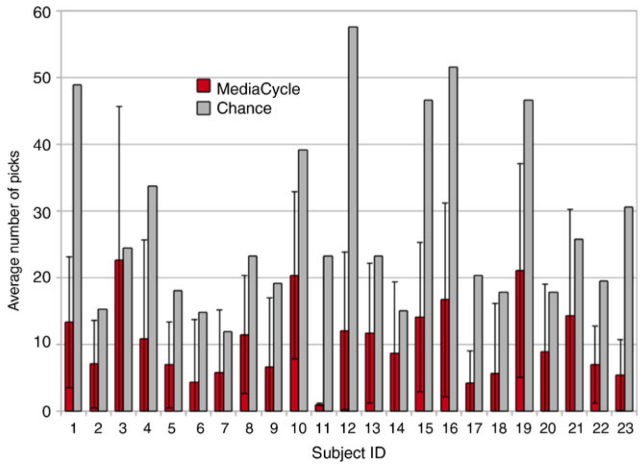

Fig. 7 Average number of picks needed to find one laughter from the same speaker

5.2.2 *Nearest neighbor laughter classification*

In this experiment, classes were built using the following laughter sublabels: vowel, nasal, breath, hiccup, hum, grunt. Speech-laughs are excluded from this study; only pure laughs are used. The sublabel "silent" is not considered since we are performing audio classification.

A laugh belongs to a class if it contains the corresponding sublabel. As mentioned in Sect. 3, each laugh can be part of several classes. For each class, each laugh was sent to MediaCycle, which returned the closest neighbor in the database ($N = 1$). The classification was considered successful if the label of the returned laugh also contained the class

[1] The lower the number of picks, the faster the search.

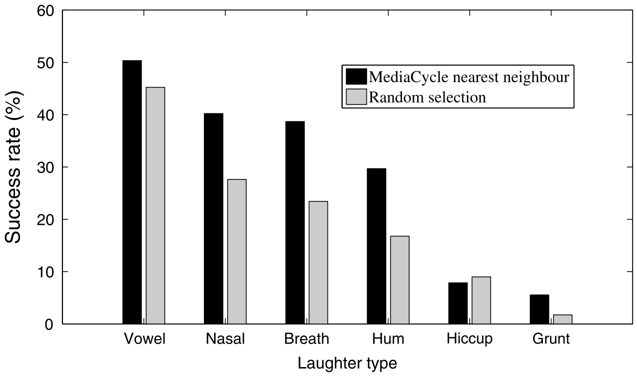

Fig. 8 Success rates achieved by MediaCycle (*black*) against chance (*gray*) for laughter classification

sublabel. Figure 8 compares the success rate of each class against chance. The performance is not outstanding. Nevertheless, MediaCycle performs better than chance. The difference is significant at a 95% confidence level ($p = 0.0138$) and these tests provide us with a first measurement in the field of laughter classification. Several paths to improve the results are suggested in Sect. 6. Due to the availability of the database and these baseline results, any improvement of the classification will be measurable. Due to the experimental conditions (one laugh can belong to several classes), it is difficult to analyze the errors. Future tests with traditional, mutually exclusive classes will be performed to better understand the errors.

5.3 Similarity and visual cues of laughter

At the moment, our similarity algorithm compares only audio features. We aimed to check whether the visual cues of laughter displayed by a virtual agent influence the perception of the similarity among laughs. For this reason, in this perceptive experiment we compare the perception of similarity between the laughter audio samples and the same samples accompanied by the animation displayed by a virtual agent. We would also like to check whether, in the case of audio samples, the participants' perception of similarity corresponds to the algorithmic labeling of similarity.

We hypothesize the following:

- *Hypothesis 1*—the visual cues of a laugh influence the user's perception of the similarities among virtual agent laughs.
- *Hypothesis 2*—audio-only laughter episodes that are similar according to our algorithm are also considered similar by participants.

5.3.1 Set-up

Three laughter input samples, one female (sample 2) and two male (samples 1 and 3), were chosen from the AVLaughterCycle database. For each input sample, $input_i$, we extracted two more samples from the database: $sampleA_i$, that was chosen among the 35% most similar audio samples, and $sampleB_i$, that was selected from the set of the 35% least similar samples. We applied this 35% threshold instead of choosing the most similar and the least similar laughs in the whole database as the latter could oversimplify the evaluation task.

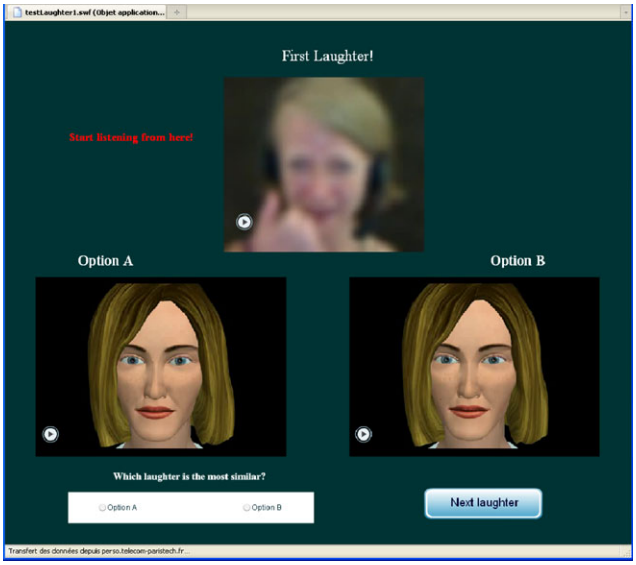
testLaughter1.swf (Objet application...
First Laughter!
Start listening from here!
Option A
Option B
Which laughter is the most similar?
○ Option A
○ Option B
Next laughter
Transfert des données depuis perso.telecom-paristech.fr...

Fig. 9 Web page of the evaluation in multimodal condition

To avoid influencing participants with characteristics that were not linked to laughter, both $sampleA_i$ and $sampleB_i$ were female (the gender of the virtual agent) and none of them belonged to the same person who emitted the corresponding input laughter $input_i$. The last criterion for selecting $sampleA_i$ and $sampleB_i$ was that they should belong to a subset of around 50 laughs for which we had a proper facial animation at the time of the experiment. All samples lasted between 2 and 15 seconds. Thus, we collected three triplets of audio samples, where each triplet a_i is

$$a_i = \{input_i, sampleA_i, sampleB_i\}.$$

In our evaluation we aimed at studying whether an animation performed by a virtual agent influences the perception of similarity among laughs. For each audio sample $sampleA_i$ and $sampleB_i$ we generated an animation with the Greta agent, using the appropriate motion capture data (see Sect. 4). The same female character was used in every animation (see Fig. 9). Thus, we collected three triplets, v_i, of multimodal (i.e. audiovisual) samples, composed of

- The video from the AVLaughterCycle database that corresponds to $input_i$.
- Two animations corresponding to $sampleA_i$ and $sampleB_i$.

The input video was slightly blurred to hide facial markers that may disturb the participants.

Table 3 Participants to the second evaluation study

Condition	Females	Males	Total
Audio	17	24	41
Mutlimodal	19	22	41
Overall	36	46	82

Participants were asked to select which laugh, *sampleA* or *sampleB*, is the most similar to the input laugh, for three triplets total. The evaluation was performed in two different conditions: audio and multimodal. In both conditions the same set of triplets was used. Each participant evaluated in only one condition. The choice of the condition as well as the order of the samples was random.

5.3.2 Procedure

82 participants (36 women, 46 men) with a mean age of 32 years took part in the study. They were mainly from France (28%), Belgium (21%), Poland (15%) and Italy (11%). 41 performed the evaluation in the audio condition (17 women, 24 men) and 41 in the multimodal condition (19 women, 22 men) (Table 3).

Participants accessed the evaluation through a web browser. One evaluation session was made of six web pages. The first one explained the evaluation study. The second page was a questionnaire where the user had to specify personal information. Each of the following three pages presented one triplet (v_i or a_i) at a time and they were displayed in a random order. The triplet depended on the condition of the study: a_i in the audio condition or v_i in the multimodal condition. Figure 9 shows an example of such a page in the multimodal condition. The participants were invited to play the input sample, $input_i$, first. Then they could play $sampleA_i$ and $sampleB_i$ as many times as they liked. $SampleA_i$ and $sampleB_i$ were displayed randomly, they could appear either on the left or on the right of the page. Before passing to the next page, the participant was obliged to select the sample estimated to be most similar to the input laughter. There was no time limit for the task. Finally, in the last page of the evaluation, the users were given the possibility to write any comment or suggestion. The participation in the study was anonymous, and the test was in English.

5.3.3 Results

With regard to the first hypothesis, that the animation of a virtual agent influences the user's perception of the similarities among laughs, we performed a Mann-Whitney test. It showed an effect of the Viewing Condition (audio vs. multimodal) on triplet 1 ($p < .05$) and on triplet 3 ($p < .05$), but not on triplet 2 ($p > .05$). Table 4 shows the results in details. The percentage of agreement with MediaCycle for all the triplets was similar in both conditions, 58.53% in audio condition and 60.16% in multimodal condition. For the first triplet in multimodal condition, all participants (100%) chose the laughter which is most similar according to MediaCycle, whereas in audio condition the percentage of agreement was 80.46%. For triplet 3, in the audio condition participants chose more often the sample that our algorithm defined as similar to the input laughter (53.65%), whereas in the multimodal condition only 34.14% of the participants agreed with MediaCycle.

Table 4 Results of the second evaluation study

	Number of subjects	Number of answers agreeing with MediaCycle	Percentage of agreement with MediaCycle
Audio			
Triplet 1	41	33	80.46%
Triplet 2	41	17	41.46%
Triplet 3	41	22	53.65%
Total	41	72	58.53%
Multimodal			
Triplet 1	41	41	100%
Triplet 2	41	19	46.34%
Triplet 3	41	14	34.14%
Total	41	74	60.16%
Overall	82	146	59.34%

No significant results were found for the second triplet. In the multimodal and audio conditions, respectively 46.34% and 41.46% of the participants agreed with our similarity algorithm.

To test our second hypothesis we analyzed the answers collected only in the audio condition and we applied the binomial test. The sample selected by our algorithm as similar was significantly more often chosen by participants for triplet 1 ($p < .05$), but it was not the case for the other two triplets ($p > .05$).

5.3.4 Discussion

With regard to the first hypothesis, results show that the visual cues play an important role in the perception of similarities among laughs. The multimodal laugh synthesized with a virtual agent from the capture motion data is perceived differently with respect to the only-audio samples. The answers to the first triplet showed that the virtual agent animation has a positive effect on user's perception of the laughs similarities, since they select more often the sample evaluated as most similar by MediaCycle. In contrast, the answers to the third triplet showed that the agent animation lowered the agreement rate between the participants and our similarity algorithm. Thus this influence does not have a uniform character. We think that not only audio features should be considered in the laughter similarity algorithm, but that video characteristics should also be taken into account.

This observation is confirmed by free comments given by some participants after the experiment. In the audio condition some participants complained that "the audio level was too low". In reality all laughter recordings were made in equal conditions but in some laughs the audio cues are discontinuous even if visual cues remain visible. This means that some laughs would be recognized and compared mainly through visual cues.

It is interesting to notice that in both cases where significant differences were observed (triplets 1 and 3), the input video was "male" while the participants had to choose between two virtual female animations and female audio samples. It may indicate some gender issues that may be studied in the future.

The second hypothesis is partially verified. Results show that in general participants more often (58.53%) chose the audio sample that was selected by our similarity algorithm. However, we obtained statistically significant results only for one triplet.

6 Conclusion

The AVLaughterCycle application has been presented in this paper. It endows a virtual agent with the capability of joining its conversational partner's laughs, by displaying a laugh response related to an input laughter. The full algorithm has been implemented and the system is operational in real time.

Several key components of the application have been evaluated. The performance of MediaCycle to retrieve similar laughs has been tested. The results form a first measurement of acoustic laughter similarity computation and show the benefits of employing MediaCycle to browse through an audio laughter database: MediaCycle provides significant improvements for grouping laughs by speaker or laughter type. In addition, subjective tests have shown that visual cues influence human perception of similarities but that users only moderately agree with our audio-only based similarity algorithm. The results of the experiments also indicate the shortcomings of our similarity algorithm. Several areas of future work are proposed here.

First, a laughter detection block could be included in SSI (until now, the input is assumed to be laughter).

Second, for the moment, the similarity analysis involves only audio timbre features. We showed that in laughter acts the visual cues are significant. Consequently they should be considered by our similarity algorithm. The feature set could also be extended to capture other important dimensions of laughter like its rhythm and structure. The weights between the different feature sets could then be tuned by the user to focus on one dimension or another. It will also be interesting to perform feature selection and see which characteristics are most relevant for specific tasks.

Third, other methods for evaluating laugh similarity could be investigated: (1) considering other distances (Mahalanobis distance, cosine similarity, etc.); (2) modeling (or resampling to a fixed length) the feature trajectories instead of taking their mean. Computing features over pulses rather than entire laughs is under development. To this end, we are annotating the database at the level of pulses, and designing an algorithm to automatically segment a laugh episode into pulses.

Fourth, other selection processes of the best laugh response can be imagined to enhance a laughter interaction, for example, the natural way of joining somebody laughing is probably not simple mimicry. Further research could be made on humans' laughter interactions to determine how one joins laughing partners.

Fifth, the automatic retargeting of the facial motion data to the animation of Greta would allow us to use the audiovisual samples immediately after the recording (i.e. when they are added to the database).

Finally, we also would like to analyze the interaction of the audio and visual data. The AVLaughterCycle database may serve to build a model of audiovisual laughter synthesis. We are also interested in the analysis of the synchronization between several acoustic and nonverbal features like breath and torso movements.

Acknowledgements This project was partly funded by the Ministry of Région Wallonne under the Numediart research program (grant N^0 716631) and by the European IP 6 project Callas. J. Tilmanne receives a Ph.D. grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium. The database participants are also gratefully acknowledged.

References

1. Bartlett MS, Littlewort GC, Frank MG, Lainscsek C, Fasel IR, Movellan JR (2006) Automatic recognition of facial actions in spontaneous expression. J Multimed 1(6):22–35
2. Becker-Asano C, Ishiguro H (2009) Laughter in social robotics—no laughing matter. In: Intl workshop on social intelligence design (SID2009), pp 287–300
3. Berk L, Tan S, Napier B, Evy W (1989) Eustress of mirthful laughter modifies natural killer cell activity. Clin Res 37(1):115A
4. Cantoche (2010) http://www.cantoche.com/
5. Cassell J, Bickmore T, Billinghurst M, Campbell L, Chang K, Vilhjálmsson H, Yan H (1999) Embodiment in conversational interfaces: Rea. In: Proceedings of the CHI'99 conference. ACM, New York, pp 520–527
6. Curio C, Breidt M, Kleiner M, Vuong QC, Giese MA, Bülthoff HH (2006) Semantic 3D motion retargeting for facial animation. In: APGV'06: Proceedings of the 3rd symposium on applied perception in graphics and visualization. ACM, New York, pp 77–84

7. DiLorenzo PC, Zordan VB, Sanders BL (2008) Laughing out loud. In: SIGGRAPH'08: ACM SIGGRAPH 2008. ACM, New York
8. Dupont S, Dubuisson T, Urbain J, Frisson C, Sebbe R, D'Alessandro N (2009) Audiocycle: browsing musical loop libraries. In: Proc of IEEE content based multimedia indexing conference (CBMI09)
9. Haptek (2010) http://www.haptek.com/
10. Heylen D, Kopp S, Marsella S, Pelachaud C, Vilhjálmsson H (2008) Why conversational agents do what they do? Functional representations for generating conversational agent behavior. In: The first functional markup language workshop, Estoril, Portugal
11. Janin A, Baron D, Edwards J, Ellis D, Gelbart D, Morgan N, Peskin B, Pfau T, Shriberg E, Stolcke A, Wooters C (2003) The ICSI meeting corpus. In: 2003 IEEE international conference on acoustics, speech, and signal processing (ICASSP), Hong-Kong
12. Knox MT, Mirghafori N (2007) Automatic laughter detection using neural networks. In: Proceedings of interspeech 2007, Antwerp, Belgium, pp 2973–2976
13. Kopp S, Jung B, Leßmann N, Wachsmuth I (2003) Max—a multimodal assistant in virtual reality construction. Künstl Intell 17(4):11–18
14. Lasarcyk E, Trouvain J (2007) Imitating conversational laughter with an articulatory speech synthesis. In: Proceedings of the interdisciplinary workshop on the phonetics of laughter, Saarbrucken, Germany, pp 43–48
15. Natural Point, Inc (2009) Optitrack—optical motion tracking solutions. http://www.naturalpoint.com/optitrack/
16. Niewiadomski R, Bevacqua E, Mancini M, Pelachaud C (2009) Greta: an interactive expressive ECA system. In: Sierra C, Castelfranchi C, Decker KS, Sichman JS (eds) 8th international joint conference on autonomous agents and multiagent systems (AAMAS 2009), IFAAMAS, Budapest, Hungary, 10–15 May 2009, vol 2, pp 1399–1400
17. Nijholt A (2002) Embodied agents: a new impetus to humor research. In: Proc Twente workshop on language technology 20 (TWLT 20), pp 101–111
18. Ostermann J (2002) Face animation in MPEG-4. In: Pandzic IS, Forchheimer R (eds) MPEG-4 facial animation—the standard implementation and applications. Wiley, New York, pp 17–55
19. Pantic M, Bartlett MS (2007) Machine analysis of facial expressions. In: Delac K, Grgic M (eds) Face recognition. I-Tech Education and Publishing, Vienna, pp 377–416
20. Parke FI (1982) Parameterized models for facial animation. IEEE Comput Graph Appl 2(9):61–68
21. Peeters G (2004) A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Tech rep, Institut de Recherche et Coordination Acoustique/Musique (IRCAM)
22. Petridis S, Pantic M (2009) Is this joke really funny? Judging the mirth by audiovisual laughter analysis. In: Proceedings of the IEEE international conference on multimedia and expo, New York, USA, pp 1444–1447
23. Pighin F, Hecker J, Lischinski D, Szeliski R, Salesin DH (1998) Synthesizing realistic facial expressions from photographs. In: SIGGRAPH'98. ACM, New York, pp 75–84
24. Ruch W, Ekman P (2001) The expressive pattern of laughter. In: Kaszniak A (ed) Emotion, qualia and consciousness. World Scientific, Singapore, pp 426–443
25. Savran A, Sankur B (2009) Automatic detection of facial actions from 3D data. In: Proceedings of the IEEE 12th international conference on computer vision workshops (ICCV workshops), Kyoto, Japan, pp 1993–2000
26. Schröder M (2003) Experimental study of affect bursts. Speech Commun 40(1–2):99–116
27. Siebert X, Dupont S, Fortemps P, Tardieu D (2009) MediaCycle: browsing and performing with sound and image libraries. In: Dutoit T, Macq B (eds) QPSR of the numediart research program, Numediart research program on digital art technologies, vol 2, pp 19–22
28. Skype Communications S à rl (2009) The skype laughter chain. http://www.skypelaughterchain.com/
29. Stoiber N, Seguier R, Breton G (2010) Facial animation retargeting and control based on a human appearance space. J Vis Comput Animat 21(1):39–54
30. Strommen E, Alexander K (1999) Emotional interfaces for interactive aardvarks: designing affect into social interfaces for children. In: Proceedings of ACM CHI'99, pp 528–535
31. Sundaram S, Narayanan S (2007) Automatic acoustic synthesis of human-like laughter. J Acoust Soc Am 121(1):527–535
32. Thórisson KR, List T, Pennock C, DiPirro J (2005) Whiteboards: scheduling blackboards for semantic routing of messages & streams. In: Thórisson KR, Vilhjálmsson H, Marsella S (eds) Proc of AAAI-05 workshop on modular construction of human-like intelligence, Pittsburgh, Pennsylvania, pp 8–15
33. Trouvain J (2003) Segmenting phonetic units in laughter. In: Proceedings of the 15th international congress of phonetic sciences, Barcelona, Spain, pp 2793–2796
34. Truong KP, van Leeuwen DA (2007) Automatic discrimination between laughter and speech. Speech Commun 49(2):144–158
35. Truong KP, van Leeuwen DA (2007) Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features. In: Proceedings of the interdisciplinary workshop on the phonetics of laughter, Saarbrucken, Germany, pp 49–53
36. Urbain J, Bevacqua E, Dutoit T, Moinet A, Niewiadomski R, Pelachaud C, Picart B, Tilmanne J, Wagner J (2010) AVLaughterCycle: an audiovisual laughing machine. In: Camurri A, Mancini M, Volpe G (eds) Proceedings of the 5th international summer workshop on multimodal interfaces (eNTERFACE'09). DIST-University of Genova, Genova
37. Urbain J, Bevacqua E, Dutoit T, Moinet A, Niewiadomski R, Pelachaud C, Picart B, Tilmanne J, Wagner J (2010) The AVLaughterCycle database. In: Proceedings of the seventh conference on international language resources and evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta
38. Vilhjálmsson H, Cantelmo N, Cassell J, Chafai NE, Kipp M, Kopp S, Mancini M, Marsella S, Marshall AN, Pelachaud C, Ruttkay Z, Thórisson KR, van Welbergen H, van der Werf R (2007) The behavior markup language: recent developments and challenges. In: 7th international conference on intelligent virtual agents, Paris, France, pp 99–111
39. Wagner J, André E, Jung F (2009) Smart sensor integration: a framework for multimodal emotion recognition in real-time. In: Affective computing and intelligent interaction (ACII 2009), Amsterdam, The Netherlands, pp 1–8
40. Zhang W, Wang Q, Tang X (2009) Performance driven face animation via non-rigid 3D tracking. In: MM '09: proceedings of the seventeen ACM international conference on multimedia. ACM, New York, pp 1027–1028
41. Zign Creations: Zign Track (2009) http://www.zigncreations.com/zigntrack.html