

AVLAUGHTERCYCLE: AN AUDIOVISUAL LAUGHING MACHINE

Jérôme Urbain¹, Elisabetta Bevacqua², Thierry Dutoit¹, Alexis Moinet¹, Radoslaw Niewiadomski²,
Catherine Pelachaud², Benjamin Picart¹, Joëlle Tilmann¹, Johannes Wagner³

¹ Laboratoire de Théorie des Circuits et Traitement du Signal (TCTS), Faculté Polytechnique de Mons (FPMs), Belgique

² CNRS - LTCI UMR 5141, Institut TELECOM - TELECOM ParisTech, Paris, France

³ Multimedia Concepts and Applications Lab (MM), Institut für Informatik, Universität Augsburg, Germany

ABSTRACT

The AVLaughterCycle project aims at developing an audiovisual laughing machine, capable of recording the laughter of a user and to respond to it with a machine-generated laughter linked with the input laughter. During the project, an audiovisual laughter database was recorded, including facial points tracking, thanks to the Smart Sensor Integration software developed by the University of Augsburg. This tool is also used to extract audio features, which are sent to a module called MediaCycle, evaluating similarities between a query input and the files in a given database. MediaCycle outputs a link to the most similar laughter, sent to Greta, an Embodied Conversational Agent, who displays the facial animation corresponding to the laughter simultaneously with the audio laughter playing.

KEYWORDS

Laughter, virtual agent, speech processing

1. INTRODUCTION

Laughter is an essential signal in human communications. It conveys information about our feelings and helps to cheer up our mood. Moreover, it is communicative, eases social contacts and has the potential to elicit emotions to its listeners. Laughter is also known to have healthy effects, and especially to be one of the best medicines against stress. Laughter therapies, “yoga” sessions or groups are emerging everywhere. Events connecting and entertaining people from all over the world through the universal signal of laughter are also successful, like the World Laughter Day or the Skype Laughter Chain [22].

In addition, the recent technological progress made the creation of a humanoid interface to computer systems possible. An Embodied Conversational Agent (ECA) is a computer-generated animated character that is 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. [3, 13]) and the industry (e.g. [2, 7]). Recent works focus on standardisation of the ECA architecture. SAIBA [28] is an international research initiative which main aim is to define a standard framework for the generation of virtual agent behaviour. It defines a number of levels of abstraction, from the computation of the agent’s communicative intention, to behaviour planning and realization. There exist several implementations of the SAIBA standard, among others SmartBody [15, 24], BMLRealizer [1], RealActor [4] and EMBR [8].

Due to the growing interest for virtual machines modeling human behaviors, a need to enable these machines to perceive and express emotions emerged. Laughter is clearly an important clue for understanding emotions and discourse events on one hand, and, on

the other hand, to manifest certain emotions and provide feedback to the conversational partners. In consequence, automatic laughter processing has gained in popularity during the last decades. However, laughter is a highly variable signal and it is hard to acoustically describe its structure. Trouvain [26] summarizes the different terminologies used in other laughter studies, as well as various categories to designate laughter types. If a few systems able to distinguish between laughter and speech have recently been built on the recognition side (e.g. [27, 12, 21]), automatic laughter synthesis is still inefficient. Interesting approaches have been explored to generate human-like laughs (e.g. [14, 23]), but perceptive tests have shown that the resulting laughs do not sound natural. They miss an important characteristic of human laughs: variability.

The AVLaughterCycle project aims at developing an audiovisual laughing machine, capable of recording the laughter of a user and to respond to it with a virtual agent’s laughter linked with the input laughter. The hope is that the initially forced laughter of the user will progressively turn into spontaneous laughter. This system will help improving emotional displays of virtual agents, and will, by itself, be an interactive application to enjoy the benefits of laughter. The virtual agent, Greta [18], will not display synthesized laughter: on the audio side, she will play an appropriately selected laughter inside an audiovisual database, and simultaneously on the visual side, she will be animated using the facial data of the selected laughter, obtained through motion capture.

The paper will be organized as follows. Chapter 2 will present the softwares used during this project: Smart Sensor Integration for recording/annotating/analyzing laughs, MediaCycle to evaluate similarities between laughs, Greta for playing the output laughter and the commercial softwares to perform motion capture, ZignTrack and OptiTrack. Chapter 3 will focus on the creation and annotation of the audiovisual laughter database, inside which utterances are selected to animate Greta. The AVLaughterCycle application process and its methods for analyzing the input laughter, selecting an answering laughter and driving Greta accordingly will be described in Chapter 4. The upcoming evaluation of the system will be discussed in Chapter 5. Finally, conclusions and future works will be presented in Chapter 6.

2. PRESENTATION OF THE TOOLS

Several important existing tools have been used in this project, as such or modified to fit our needs. In this Chapter, these tools will be presented separately. Their integration in the whole project will be described in Chapters 3 and 4.

2.1. Smart Sensor Integration (SSI)

Smart Sensor Integration (SSI) [29] is a software designed by the University of Augsburg to deal with multimodal signal recording and processing. It provides a Graphical User Interface (GUI) to start and stop a recording. Afterwards the recorded data can be visualized and annotated. Given an annotation it is possible to automatically extract features and train a model. The different modalities are automatically synchronized.

The GUI includes a dedicated space to present stimuli, which is useful for database recordings. The stimuli are presented via HTML pages automatically managed by the SSI application. Browsing through successive HTML pages supports two different modes: clicks by users or automatic page switch either after a predefined time or after a certain number of events detected in the recorded data (e.g. a certain number of laughs). As well as recording the data, SSI stores the stimuli sequence.

SSI integrates signal processing libraries. Recorded signals can be analyzed in real-time or offline. Features are defined in a Dynamic Link Library used by the SSI GUI. Every kind of signal processing algorithm can be implemented there. One can also define “triggers”: functions that decide whether or not a signal segment should be further processed; for example, employing a Voice Activity Detection on an audio input to only process parts of the signal(s) where there is vocal activity.

Using the triggers, SSI pre-segments the data. Pre-labels can be assigned to the segments via the HTML stimuli manager: when, for example, a funny stimuli is presented and laughs are expected, we can specify to assign a “laughter” label to every segment that respects the trigger conditions. The SSI GUI enables to further annotate the recordings by adding/removing segments, refining their boundaries or change their labels.

Once data is annotated, classifiers may be trained to model the feature distributions of the different classes. SSI provides classifiers implemented in the Torch3D library: Hidden-Markov Models (HMMs), Gaussian Mixture Models (GMMs), k-Nearest Neighbours (kNNs), etc. The trained classifiers can be used to label new data (in real-time or not) and can also serve as triggers in the SSI processing chain.

To summarize, SSI provides convenient methods for multimodal database recordings, annotation, classification and processing. The different aspects are used in AVLaughterCycle and SSI is integrated in the system architecture. This first integration already provides satisfying results, but will also enable us to improve the AVLaughterCycle application by exploiting more of the SSI processing and classification capabilities in the future.

2.2. MediaCycle

MediaCycle is a software developed at the University of Mons and the Université Catholique de Louvain for browsing through multimedia libraries, in the Framework of numediart Belgian R&D program. It started by considering acoustic similarities only, in a project called AudioCycle [6], designed to ease the navigation inside a large audio loops libraries. The software computes acoustic features - characterizing musical properties of rhythm, melody and timbre - for each file in an audio loop database and then evaluates the similarities between loops through the distances between their feature vectors. A Graphical User Interface has also been designed to visualize the database: loops are grouped into clusters through a K-means algorithm; a reference loop is randomly selected and other loops are positioned around it according to their cluster belonging and similarities with the reference loop. This is illustrated

in Figure 1. Tools to easily browse through the library are available such as playing any combination of loops, which are synchronized, reorganizing the database by selecting a new reference loop or changing the features weights, splitting again one cluster, etc.

AudioCycle has been extended to deal with visual content in a project called MediaCycle where image features were added. Methods for computing similarities between videos are also progressively implemented, as well as dedicated methods to process laughter, which received a particular attention in another subproject called LaughterCycle. The system can also be queried by laughing: it then places the incoming laughter in the database space and outputs the N most similar utterances.

In this project, the visual database organization provided by MediaCycle will not be used since the visual output is performed by Greta. Only the MediaCycle engines for organizing a database and computing similarities between objects will be integrated in AVLaughterCycle.

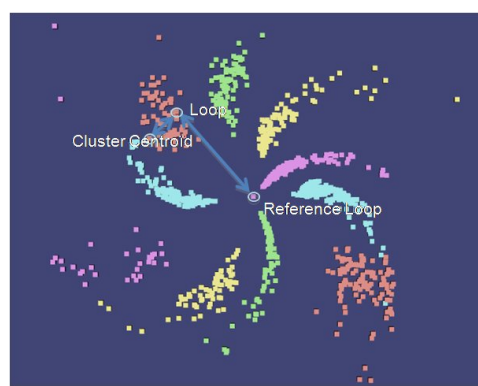


Figure 1: AudioCycle Database visualization

2.3. Motion Capture

Motion capture (often referred to as “mocap”) consists in recording a real motion by transcribing it under a mathematical form usable by a computer. This is achieved by tracking a number of key points through space across time, and combining them to obtain a tridimensional unified representation of the performance [16]. Several techniques can be used for motion capture, but when it comes to facial motion capture, only techniques that do not need intrusive equipment nor large markers can be considered.

Facial motion capture can be divided into two main types: marker and markerless techniques. While markerless techniques have the huge advantage that they can usually be performed on facial videos recorded without any specific setup, they are nowadays not robust enough to ensure automatic and reliable capture of the small variations of facial expression during laughter for instance. Using markers placed on the face eases the tracking and increases its robustness. However, the recording must then be performed in an unnatural setting, and very often with dedicated equipment and quite heavy setups.

In this project it was chosen to record our database using marker based tracking, in order to obtain data with as much precision and information as possible. Two different commercial motion capture tools, ZignTrack and OptiTrack, have been tested and used. They are presented below.

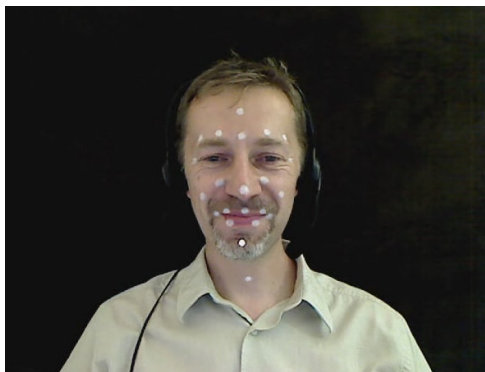


Figure 2: ZignTrack - 22 markers on the subject's face

- **ZignTrack** [5] is the first software we used. It captures 3D facial motion with one single camera. It requires 22 facial features, marked with simple stickers or make-up (no special markers or infrared equipment required), as illustrated in Figure 2. They must be not too small, to be obvious on the video, nor too big, to track accurately the face features. ZignTrack handles head rotation, jaw/lip-syncing, eyebrows, eyelids sneer and cheek movements. As the subject is recorded with one single camera placed in front of him, the markers are actually tracked in a two dimensional space only, and the 3D points are extrapolated by the software using a fixed face template. The transformations linked to the head displacements and rotations are not all perfectly handled by the software, as we noticed for instance that the face was “shrinking” or “inflating” when there were up and down rotations of the head.

Once the recording is over, the video is imported into ZignTrack, the face markers are manually linked to a template on the first video frame, and the tracking of those markers is then automatically performed. In case of tracking errors, manual tuning of each marker position can be performed to adjust/correct the automatic tracking.

Once the tracking is done, ZignTrack extrapolates the 3D positions of the tracked face points. The resulting motion capture data can then be exported to several motion capture formats: BVH, TRC, Poser pz2 and Animation:Master action files. The ZignTrack software costs around 130 euros.

- **OptiTrack** [17] is an Optical Motion Capture solution developed by Natural Point. Our seven-cameras face motion tracking desktop setup is shown in Figure 3 (the positions of the 7 cameras are marked by a red circle).

The seven synchronized infrared cameras are placed in a semi-circular way: six for face motion capture and a middle additional one for scene A/V recording (recording synchronized audio and video tracks for each take). For each of them, a grayscale CMOS imager captures up to 100 frames per second.

OptiTrack requires at least 23 face markers and 4 head markers on the actor (Figure 4). These markers are infrared reflectors stuck on the skin, smaller than the white make-up dots of the ZignTrack device. Therefore they provide a very accurate and robust (versus head movements) face tracking using OptiTrack Arena Facial Expression software. The results can be exported to various formats such as BVH, C3D.

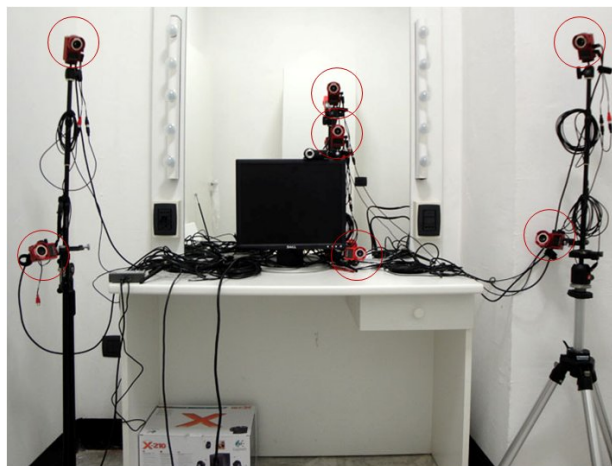


Figure 3: Desktop setup for facial motion tracking using OptiTrack



Figure 4: OptiTrack - 23 face markers and 4 head markers on the subject

The whole OptiTrack desktop setup is much more expensive than ZignTrack, and costs around 5000 euros. OptiTrack is not limited to face motion tracking but can be upgraded to cover full body motion tracking, by adding more cameras and sensors.

2.4. The 3D humanoid agent: Greta

Greta [18] (Figure 5) is a 3D humanoid agent developed by Telecom-ParisTech. She is able to communicate with the user using verbal and nonverbal channels like gaze, facial expressions and gestures. It follows the SAIBA framework [28] that defines a modular structure, functionalities and communication protocols for Embodied Conversational Agents (ECAs). Moreover, Greta follows the MPEG4 [19] standard of animation. Greta uses the FML-APML XML-language [9] to specify the agent's communicative intentions (e.g., its beliefs, emotions) that go along with what the agent wants to say. The communicative intentions of the listener are generated by the Listener Intent Planner while the intentions of the speaker are defined at the moment manually in an FML-APML input file. The Behavior Planner module receives as input the agent's communicative intentions written in FML-APML and generates as output a list of signals in BML language. BML specifies the verbal and nonverbal behaviors of the agent [28]. 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 that generates the MPEG4 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 a whiteboard. For this purpose we use Psyclone messaging system [25] which allows modules and applications to interact through TCP/IP. The system has a very low latency time that makes it suitable for interactive applications.



Figure 5: Greta, the 3D humanoid agent used in AVLaughterCycle

3. CREATION OF AN AV LAUGHTER CORPUS

The first step of the AVLaughterCycle project is the recording of an audiovisual (AV) database consisting of humans laughing. Participants of the eNTerFACE09 workshop who wanted to contribute to this experiment were invited to laugh in front of the set of cameras. This Chapter is intended to describe the laughter database. It is divided in 7 sections presenting: the selection of stimuli and instructions given to the participants (Section 3.1), the settings for audiovisual recording (Section 3.2) and facial motion capture (Sections 3.3 and 3.4), the corpus annotation protocol (Section 3.5), the participants (Section 3.6) and, finally, the database contents (Section 3.7).

3.1. Elicitation method: selection of stimuli, protocol of DB recording

It is known that there is a difference between the expressions of real and acted emotions (e.g. [30]). To collect a corpus representative of humans' natural behaviours, one should try to capture the data in a natural environment, the subjects being unaware of the database collection until the end of the recording. Laughter being an emotional signal, it is affected by the same phenomenon: one cannot expect natural laughter utterances by simply asking subjects to laugh. To find spontaneous laughter utterances, it is popular to take the laughs recorded while collecting data for another purpose. For example, [27, 12] and [11] use the ICSI Meeting Corpus [10], recorded for studying speech in general by placing

microphones in meeting rooms. Apart from speech, this corpus contains a significant number of laughs, which are assumed spontaneous since they occur in regular conversations (even though the participants knew there were microphones). When for some reason natural data cannot be used, it is common to try to induce laughter - and not tell laughter is the object of the study - rather than asking to laugh. One way to achieve it is to display a funny movie.

In our case, both audio recording and accurate facial motion tracking were needed. To our knowledge, there existed no laughter database providing these 2 signals. Due to the markers required for facial motion tracking and the fact that subjects should stay in the camera(s) space, a natural laughter recording was impossible. To push the participants towards spontaneous laughter, even though they knew that they were being recorded, a 13-minutes funny movie was created by the concatenation of short videos found on the internet. The participants were asked to relax, watch the video and enjoy it. They could close their eyes, move a bit their head but should keep it towards the camera during the whole recording. Moreover, they could not put anything between their head and the webcam (e.g. hands), else the face tracking is lost. Except these two limitations, they could act freely, talk, laugh, cry, shake their head, etc. as they would do if they were at home. At the end of the experiment, subjects were instructed to perform one acted laughter, pretending they had just heard/seen something hilarious.

3.2. Audiovisual (AV) laughter database recording

The AV laughter database was recorded on site (Casa Paganini, Genova, Italy), using one webcam (ZignTrack) plus seven infrared cameras (OptiTrack) for video recording, a headset for audio recording (16 kHz, 16 bits/sample) and stimuli listening, and University of Augsburg's Smart Sensor Integration tool (SSI) for stimuli playing and audio/video recording synchronization (and later for recordings annotation). All these components have already been described in Chapter 2.

3.3. Facial motion capture using ZignTrack

The webcam used has a 640x480 resolution, stores in RGB 24 bits and captures 25 frames per second (FPS). The 22 marker dots were simply made of white make-up (Figure 2). Another attempt was performed using red markers but the face tracking failed because of the poor contrast between them and the skin colour.

ZignTrack is a cheap facial motion capture software, working quite well with markers that stay visible during the whole recording and with slow head movements (fuzzy effect due to the 25 FPS limitation in case of fast head movements). If at least one of these constraints is not respected, the tracking fails, requiring heavy manual corrections. This is the reason why we turned towards a more sophisticated (and more expensive) system, OptiTrack.

3.4. Facial Motion Capture using OptiTrack

During this work, the seven-cameras Face Motion Tracking desktop setup shown in Figure 3 was used. Compared to the basic OptiTrack system, our setup also includes the webcam in parallel in order to be able to use the SSI software and keep all its previously explained advantages. A post-processing is carried out to synchronize the webcam with the OptiTrack cameras.

The OptiTrack Arena Facial Expression software performed very well and provided a more robust tracking than ZignTrack. Indeed, even if some markers are lost during the tracking (e.g. too large head rotations), they are nearly always recovered after a short period of time, thanks to the number of cameras and points of view (six) as well as to infrared (versus visible spectrum) acquisition performance. In addition, each individual marker can also be manually tuned to adjust/correct the automatic tracking if it does not recover by itself.

However, it seems that recordings longer than 5 minutes are not always completely saved (the end is sometimes missing). We thus decided to reduce the 13-minutes funny movie to 10 minutes, and to split it into 3 parts of around 3 minutes each.

3.5. Database Annotation

The recorded data have been annotated using SSI. A hierarchical annotation protocol was designed: segments receive the label of one main class (laughter, breath, verbal, clap, silence or trash) and “sublabels” can be concatenated to give further details about the segment. The main objective of the sublabels is to distinguish between different kinds of laughs, but still being able to rapidly group subclasses when needed, for example when only the main classes are relevant. Laughter sublabels characterize both:

- the laughter temporal structure: following the three segmentation levels presented by Trouvain [26]. 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.
- the laughter acoustic contents: through labels referring to the type of sound: voiced, breathy, nasal, grunt-like, hum-like, “hiccup-like”, speech-laughs or laughs that are mostly visual (quasi-silencious).

While only one main class can be assigned to a segment, sublabels can be combined, for example to indicate that the laughter episode contains several bouts and that we can find hiccup-like and voiced ‘a’ parts in it. To cope with exceptional classes conflicts that might influence the classes models when training a classifier - for example when we can hear a phone ringing in the middle of a laughter episode - a ‘discard’ main class has been added.

The annotation primarily relies on the audio, but the video is also looked at, to find possible neutral facial expressions at the episode boundaries or annotate visual-only laughs. In addition, laughs are often concluded by an audible inspiration, sometimes several seconds after the laughter main part. When such an inhalation, obviously due to the preceding laughter, can be found after the laughter main audible part, it is included in the laughter segment.

The annotated laughs form our laughter database, inside which an answering laughter is selected when AVLaughterCycle is queried (see Chapter 4).

3.6. Participants

24 subjects participated in the database recordings: 8 (3 females, 5 males) with the ZignTrack setting and 16 (6 females, 10 males) with the OptiTrack setting. They came from various countries: Belgium, France, Italy, UK, Greece, Turkey, Kazakhstan, India, Canada, USA and South Korea. The male average age was 28 (standard deviation: 7.1) and the female average age was 30 (sd: 7.8), which correspond to a global average age of 29 (sd: 7.3). All

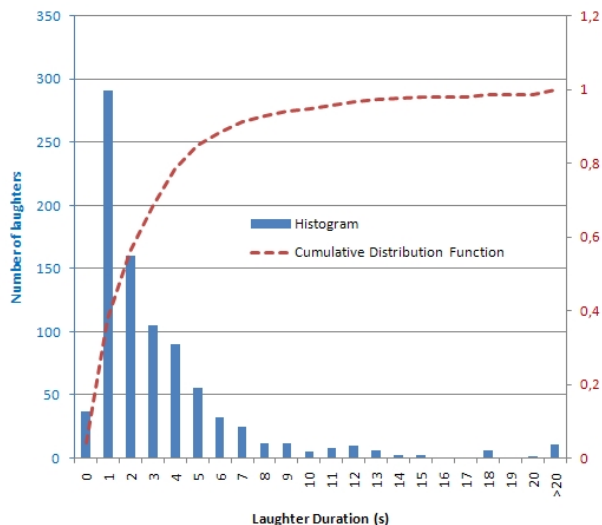


Figure 6: Histogram and cumulative distribution function of the laughs durations

the participants gave written consent to use their data for research purposes.

3.7. Database contents

Annotation (Section 3.5) is still under way, but from the 20 files that are already fully annotated, preliminary analyses of the corpus contents can be performed: subjects spend, in average, 23.5% of the recording laughing, which is a huge amount of time. The number of laughter episodes per participant stands around 43.6, with extreme values of 17 and 82, for a total of 871 episodes in these 20 files. The average duration of a laughter episode is 3.6s (standard deviation: 5.5s). A histogram of the laughs durations and their cumulative distribution function is presented in Figure 6. The large majority (82%) of the laughter episodes lasts less than 5s, but longer episodes should not be neglected as they represent 53.5% of the total laughs duration and, above all, are the most striking ones. The longest giggle in the analyzed database lasts 82s.

4. CORPUS BASED AUDIOVISUAL LAUGHTER SYNTHESIS

The communications between the different modules are illustrated in Figure 7. The dashed arrows refer to the database building process. In this Chapter, the AVLaughterCycle application process will be described (solid arrows). Users can query the AVLaughterCycle systems 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 (with a trigger to delimit laughter segment boundaries). In both cases, when the audio laughter segment is available, SSI computes the segment features (see Section 4.1) and sends them to MediaCycle. MediaCycle compares these features with the database samples and outputs the most similar example, as reported in Section 4.2. This output is sent to Greta who plays the audio sound synchronously with the corresponding facial animation. To do so, Greta had to be slightly modified. This will be explained in Sections 4.3 and 4.4.

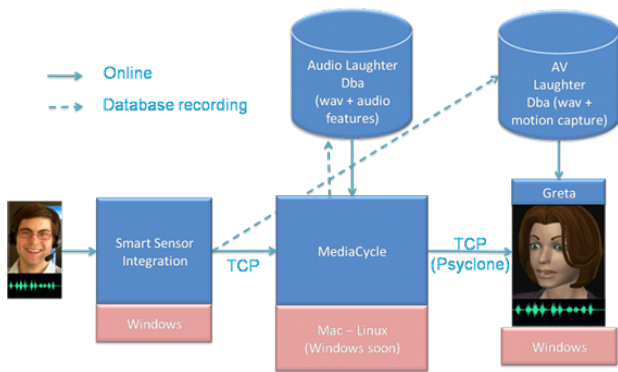


Figure 7: Flow chart of the AVLaughterCycle application

4.1. Laughter audio similarity analysis

The labeled laughter 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' set [20] and implemented in a C++ library. The features can be extracted directly in SSI, where the MediaCycle audio feature extraction library has been integrated, and then sent to MediaCycle. In the current demo version, we used the following spectral features:

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

In addition, 2 temporal features were used: the energy and the zero-crossing rate. In total, 60 features were extracted for each frame of 213ms, with 160ms overlap. The similarity estimation requires comparing audio segments of different lengths, hence different numbers of frames. To obtain a constant features vector size, we 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 already provides promising results. It had been successfully used in other similarity computation contexts [6] and was assumed applicable to laughter timbre characterization.

Euclidian distance between feature vectors is used to compute the dissimilarity between laughter episodes. For the moment, the similarity analysis involves only audio timbre features. It is planned to include audio rhythmic features, for which most of the algorithms are available in MediaCycle but processing of some exceptional cases should be improved to avoid unexpected behaviours in uncontrolled conditions (real-time use). Furthermore, visual features could be added to the similarity measures in the future, to take an audiovisual similarity decision. MediaCycle provides image/video feature extraction methods that could be used.

Furthermore, the weighting of the different feature sets can be defined and modified in real-time by the user, to put the focus more on audio rhythm or video features, for instance.

4.2. Answering laughter selection

When AVLaughterCycle is queried, the input laughter is analyzed and his feature vector is computed. This vector is used to select a corresponding answering laughter inside the laughter database organized by MediaCycle. In this project, it was decided to output the closest (i.e. most similar according to our feature set) laughter from the input laughter. Doing this way, the system can be employed to search inside the database for a specific kind of laughter. However, other selection processes can be imagined to enhance a laughter interaction: the natural way of joining somebody laughing is probably not to mimic him. Further research could be made on humans' laughter interactions to determine how we join laughing partners, model it and integrate that into MediaCycle's best answering laughter selection.

4.3. Visual Replay

The data from the motion capture softwares contain, for each frame, the position of each marker in the 3D space. Thus values for face are influenced by head rotations and body movements. First of all, these movements, used to animate Greta's general posture (with BAPs), were separated from the facial movements. Noise caused by the technical flaw of the capturing hardware was also removed. Such a data was used to animate Greta agent that uses MPEG-4 standard of animation. In this standard, the face model is animated by using 66 points called FAPs. Each of them deforms one region of the face in one direction (i.e. horizontal or vertical). Two different mappings were created to map the motion capture data, coming from ZignTrack or OptiTrack, to FAPs standard. Unfortunately many FAP points do not correspond to any marker. Thus it was not possible to use simple one-to-one mapping. For several FAPs, linear combinations of several markers values and weights were defined. Last but not least the motion capture data stored in the laughter database correspond to different face geometries of different subjects while our virtual agent uses only one face geometry model. Thus the captured movements of different persons (e.g. widely open mouth) had to be adapted to Greta's model. For this propose the mappings can be parameterized to cover the inter-personal variability.

4.4. GRETA is playing the analyzed facial signals from triggered AV laughter selected from the database

For the AVLaughterCycle application, Greta underwent certain modifications. Greta's behaviour is usually defined in BML language. It specifies the verbal and nonverbal behaviours of the agent. Each BML tag corresponds to a behaviour the agent produces on one modality: head, torso, face, gaze, gesture, speech. Single nonverbal behaviours are defined using high level symbolic representation. On the other side our laughter database contains the precise, frame-by-frame descriptions of partial animations (i.e. only the face) in FAP format.

In this project, the default BML syntax was extended to allow mixing (high level) BML commands with (low level) FAPs description and we modified Greta's animation engine to be able to generate a smooth animation for such a mixed content. Consequently Greta may display a laughter animation using the data

from the laughter corpus which is accompanied by an audio file and other nonverbal signals that might be specified in BML language (like gestures, or other facial expressions). Greta was integrated in the AVLaughterCycle architecture using Psychone and BML commands. It allows for immediate visualization of audiovisual response to user's detected laughter.

5. EVALUATION OF LAUGHTER SYNTHESIS BASED ON LAUGHTER SIMILARITY

In order to assess the validity and efficiency of the developed laughter analysis and synthesis chain, an evaluation study will be carried out. The objective is to measure the similarity of the audio answer as well as the improvement brought by the visual display. The protocol will be the following.

To assess the similarity algorithm, subjects will be presented laughs by pairs and will be asked to rate their similarities on a Likert scale (1 to 7). The laughter pairs will be formed the following way: an input laughter will be selected to query the MediaCycle device, which will output 3 laughs: the most similar one, the least similar one and an average distance laughter; three pairs will then be constituted, each one gathering the input laughter (unmodified) and one of the MediaCycle outputted laughs. The same process will be repeated with a number (at least 10) of input laughs and the pairs will be randomly ordered for each subject.

To measure the improvement brought by the visual display, three different conditions will be tested: using audio only, using video only and combining both modalities. When video is used, it consists in a Greta animation driven by the corresponding laughter facial animation. Due to the number of pairs to compare, it is planned to have 3 sets of subjects and assigning each group to only one modality. At least 10 subjects will be needed in each group.

Mixed Anova tests will be performed to evaluate whether the most similar output of MediaCycle is indeed perceived as closer from the input laughter than the 2 other inputs (least similar and average distance) as well as looking to the differences across conditions.

6. CONCLUSION

AVLaughterCycle, a software for real-time recording of laughter and playing of an acoustically similar laughter by an Embodied Conversational Agent, has been presented. The main deliverables of the project are the large audiovisual laughter database, that will be released fully annotated, and the integration of several different modules into one single processing chain to implement all the steps from laughter recording to similar output playing. Several issues were encountered during the project. We can cite communication bugs between SSI and MediaCycle, difficulties of automatic Facial Tracking during laughter or mapping from Motion Capture Data to Greta animation. Solutions were proposed for these issues during the project. Demos shown encouraging results but also revealed some lack of robustness in the similarity computation, which is the main focus for future developments. An evaluation of the device will be carried out in order to numerically characterize its efficiency. Means to automatize or ease the mapping between the motion capture data and Greta's animation will also be investigated. Other suggested future works include voice/character conversion (to avoid having one single agent laughing with different voices), integration of visual features in the similarity computation and building models to not only reproduce but synthesize laughter,

as well as to imitate how humans respond to conversational partners' laughs.

7. ACKNOWLEDGEMENTS

The authors would like to thank all the eNTERFACE'09 attendants who participated in the creation of the database.

This project was partly funded by the numediart research project, funded by Région Wallonne, Belgium (grant N°716631), and by the European IP 6 project CALLAS.

Joëlle Tilmanne receives a PhD grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium.

8. REFERENCES

8.1. Scientific references

- [3] J. Cassell et al. "Embodiment in conversational interfaces: Rea". In: *In CHI* (1999). P.: 97.
- [4] Aleksandra Cerekovic, Tomislav Pejisa, and Igor S. Pandzic. "RealActor: Character Animation and Multimodal Behavior Realization System". In: *Intelligent Virtual Agents, 9th International Conference, IVA 2009, Amsterdam, The Netherlands, September 14-16, 2009, Proceedings*. Ed. by Zsófia Ruttkay et al. Vol. 5773. Lecture Notes in Computer Science. Springer, 2009. Pp. 486–487. P.: 97.
- [6] S. Dupont et al. "AudioCycle : Browsing Musical Loop Libraries". In: *Proc. of IEEE Content Based Multimedia Indexing Conference (CBMI09)*. Chania, Greece 2009. Pp.: 98, 102.
- [8] Alexis Heloir and Michael Kipp. "EMBR - A Real-time Animation Engine for Interactive Embodied Agents". In: *Intelligent Virtual Agents, 9th International Conference, IVA 2009, Amsterdam, The Netherlands, September 14-16, 2009, Proceedings*. Ed. by Zsófia Ruttkay et al. Vol. 5773. Lecture Notes in Computer Science. Springer, 2009. Pp. 393–404. P.: 97.
- [9] D. Heylen et al. "Why Conversational Agents do what they do? Functional Representations for Generating Conversational Agent Behavior". In: *The First Functional Markup Language Workshop*. Estoril, Portugal 2008. P.: 99.
- [10] A. Janin et al. "The ICSI Meeting Corpus". In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Hong-Kong 2003. P.: 100.
- [11] L. Kennedy and D. Ellis. "Laughter detection in meetings". In: *NIST ICASSP 2004 Meeting Recognition Workshop*. Montreal 2004. P.: 100.
- [12] M. T. Knox and N. Mirghafori. "Automatic Laughter Detection Using Neural Networks". In: *Proceedings of Interspeech 2007*. Antwerp, Belgium 2007. Pp. 2973–2976. Pp.: 97, 100.
- [13] S. Kopp et al. "Max - A Multimodal Assistant in Virtual Reality Construction". In: *KI 17.4* (2003). Pp. 11–18. P.: 97.
- [14] E. Lasarczyk and J. Trouvain. "Imitating conversational laughter with an articulatory speech synthesis". In: *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*. Saarbrücken, Germany 2007. Pp. 43–48. P.: 97.

- [15] J. Lee and S. Marsella. “Nonverbal Behavior Generator for Embodied Conversational Agents”. In: *Proceedings of 6th International Conference on Intelligent Virtual Agents*. Vol. 4133. LNCS. Marina Del Rey, CA, USA: Springer, 2006. Pp. 243–255. P.: 97.
- [16] A. Menache. *Understanding motion Capture for Computer Animation and Video Games*. San Francisco, CA, USA: Morgan Kauffman Publishers Inc., 1999. P.: 98.
- [18] Radoslaw Niewiadomski et al. “Greta: an interactive expressive ECA system”. In: *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, May 10-15, 2009, Volume 2*. Ed. by Carles Sierra et al. IFAAMAS, 2009. Pp. 1399–1400. Pp.: 97, 99.
- [19] J. Ostermann. “MPEG-4 Facial Animation - The Standard Implementation and Applications”. In: Wiley, England 2002. Chap. Face Animation in MPEG-4, pp. 17–55. P.: 99.
- [20] G. Peeters. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. Tech. rep. 2004. P.: 102.
- [21] Stavros Petridis and Maja Pantic. “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 2009. Pp. 1444–1447. P.: 97.
- [23] S. Sundaram and S. Narayanan. “Automatic acoustic synthesis of human-like laughter”. In: *Journal of the Acoustical Society of America*. Vol. 121. 1. 2007. Pp. 527–535. P.: 97.
- [24] M. Thiébaux et al. “SmartBody: behavior realization for embodied conversational agents.” In: *Proceedings of 7th Conference on Autonomous Agents and Multi-Agent Systems*. 2008. Pp. 151–158. P.: 97.
- [25] K. R. Thórisson et al. “Whiteboards: Scheduling blackboards for interactive robots”. In: *Twentieth National Conference on Artificial Intelligence*. 2005. P.: 100.
- [26] Jurgen Trouvain. “Segmenting Phonetic Units in Laughter”. In: *Proceedings of the 15th International Congress of Phonetic Sciences*. Barcelona, Spain 2003. Pp. 2793–2796. Pp.: 97, 101.
- [27] K. P. Truong and D. A. van Leeuwen. “Automatic discrimination between laughter and speech”. In: *Speech Communication* 49 (2007). Pp. 144–158. Pp.: 97, 100.
- [28] H. Vilhjalmsson et al. “The Behavior Markup Language: Recent developments and challenges”. In: *7th International Conference on Intelligent Virtual Agents*. Paris, France 2007. Pp.: 97, 99.
- [29] Johannes Wagner, Elisabeth André, and Frank Jung. “Smart sensor integration: A framework for multimodal emotion recognition in real-time”. In: *Affective Computing and Intelligent Interaction (ACII 2009)*. 2009. P.: 98.
- [30] Janneke Wilting, Emiel Kraemer, and Marc Swerts. “Real vs. acted emotional speech”. In: *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 ICSLP)*. Pittsburgh, USA 2009. Pp. 805–808. P.: 100.

8.2. Software, hardware and technologies

- [1] B.P. Árnason and A. Þorsteinsson. “The CADIA BML Realizer”. URL: <http://cadia.ru.is/projects/bmlr/>. P.: 97.
- [2] Cantoche. URL: <http://www.cantoche.com>. P.: 97.
- [5] Zign Creations. “Zign Track - The affordable facial motion capture solution”. 2009. URL: <http://www.zigncreations.com/zigntrack.html>. P.: 99.
- [7] Haptek. URL: <http://www.haptek.com>. P.: 97.
- [17] Natural Point, Inc. “OptiTrack - Optical motion tracking solutions”. 2009. URL: <http://www.naturalpoint.com/optitrack/>. P.: 99.
- [22] Skype Communications. “The Skype Laughter Chain”. 2009. URL: <http://www.skypelaughterchain.com>. P.: 97.