The AVLaughterCycle database

Jérôme Urbain[1], Elisabetta Bevacqua[2], Thierry Dutoit[1], Alexis Moinet[1], Radoslaw Niewadomski[2], Catherine Pelachaud[2], Benjamin Picart[1], Joëlle Tilmanne[1] and Johannes Wagner[3]

[1] TCTS Lab, Faculté Polytechnique, Université de Mons, Boulevard Dolez 31, 7000 Mons, Belgium
[2] CNRS – LTCI UMR 5141, Institu TELECOM – TELECOM ParisTech, 46 rue Barrault, 75013 Paris, France
[3] Institut für Informatik, Universität Augsburg, Universitätsstr. 6a, 86159 Augsburg, Germany

As part of the AVLaughterCycle project carried out during the eNTERFACE'09 Workshop held in Genova, an audiovisual laughter database has been recorded. The aim of this database is to provide a broad corpus for studying the acoustics of laughter, the facial movements involved, and the synchronization between these two signals. During the Workshop, the laughter database has been used to drive the facial movements of a 3D humanoid virtual character, Greta [1], simultaneously with the audio laughter signal. In this paper the database collection protocol will be detailed.

1) 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 (see Section 4). They came from various countries: Belgium, France, Italy, UK, Greece, Turkey, Kazakhstan, India, Canada, USA and South Corea. The female, male and overall average ages were respectively 30 (standard deviation: 7.8), 28 (sd: 7.1) and 29 (sd: 7.3). All the participants gave written consent to use their data for research purposes.

2) Stimuli
It is known that there is a difference between the expressions of real and acted emotions (e.g. [2]). 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 laughters recorded while collecting data for another purpose. For example, [3, 4, 5]use the ICSI Meeting Corpus [6], recorded for studying speech in general by placing microphones in meeting rooms. Apart from speech, this corpus contains a significant number of laughters, 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 beforehand that laughter is the object of the study - rather than asking to laugh. One way to achieve it is to display a funny movie [10].

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, a natural laughter recording was impossible. To push the participants towards spontaneous laughter, a 13-minutes funny movie was created by the concatenation of short videos found on the internet.

3) Database recording protocol

Participants were invited to sit in front of a computer screen. They wore a headset microphone for audio recording and stimuli listening. The funny movie was displayed on the screen. A webcam was placed on top of the screen, recording 25 frames per second (FPS) with a 640x480 resolution, stored in RGB 24 bits. The audio sampling frequency was set to 16kHz, stored in PCM 16 bits. The material for facial motion capture will be presented in Section 4.

The database was recorded through University of Augsburg's Smart Sensor Integration (SSI) [7]. This software enables the synchronization between the different input signals (here, microphone and webcam), handles the stimuli display and can directly process the signals to segment and label interesting parts. SSI was also used for the database annotation (Section 5).

Participants received few instructions. They were asked to relax, watch the video and enjoy it. They could close their eyes, move a bit their head but should try to keep it towards the screen 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. Once the instructions were clear, participants were left alone until the end of the experiment. At the end of the movie, subjects were instructed to perform one acted laughter, pretending they had just heard/seen something hilarious.

4) Facial Motion capture

Since markerless facial motion tracking is nowadays not reliable enough to capture the small variations of facial expression during laughter, we turned towards techniques using markers placed on the subject's face. Two different systems have been successively used, ZignTrack and OptiTrack:

a) ZignTrack

ZignTrack [8] uses one single camera to realize the 3D tracking, which is indeed an extrapolation from a 2D image, using a fixed face template. Facial features are marked with simple stickers or make-up (Figure 1). ZignTrack presents the advantages of being cheap and requiring few material, but has also several drawbacks: the extrapolation from 2D causes head distortions, the tracking of the facial points fails when there are rapid movements and the tracking is unable to recover after an erroneous frame. To obtain the accurate facial motion, a lot of manual corrections are then needed. For these reasons, we turned towards a more professional device, OptiTrack, after the first 8 recordings.

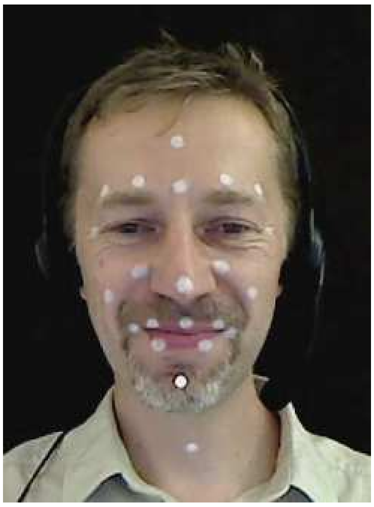

Figure 1: Markers drawn for facial motion tracking using ZignTrack

b) OptiTrack

OptiTrack [9] uses 7 synchronized infrared cameras placed in a semi-circular way: 6 for facial motion capture and an additional one for scene audiovisual recording. Each camera contains a grayscale CMOS imager capturing up to 100FPS. Infrared reflectors need to be stuck on the skin (Figure 2). For the 16 recordings performed with the OptiTrack device, the 7 infrared cameras were added to the previous setting. Participants were asked to clap their hands in order to synchronize the facial motion tracking with the audio and webcam signals. OptiTrack provided high quality tracking with few manual corrections required. However, the infrared camera data acquisition sometimes stopped after around 5 minutes. To make sure the data of the whole experiment would be usable, it was thus decided to reduce the video to 10 minutes and to split it in 3 parts slightly longer than 3 minutes. Each video part was recorded separately and the acquisition system was started and synchronized again for each part.

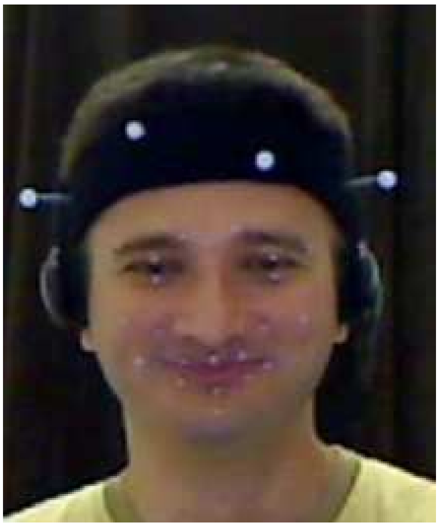

Figure 2: Infrared markers placed for facial motion tracking using OptiTrack

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 or trash; silence being the default class) 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 laughters, 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 [10]. 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 laughters 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 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 laughters. In addition, laughters 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.

6) Database contents

Annotation 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 number of occurrences of the main classes and the laughter sublabels are respectively presented in Tables 1 and 2. It is important to note that the sublabels could be combined to indicate that the laughter episode presents several different contents. This explains why the total number of sublabels is larger than the number of occurrences in the laughter class.

Main class	**Occurrences**
Laughter	912
Trash	201
Verbal	135
Clap	65
Breath	34
Discard	28

Table 1: occurrences of the main classes

Category	Laughter sublabel	Occurrences
Duration	Monosyllabic	157
	One bout (several syllables)	607
	More than one bout	143
Acoustic content	Voiced	406
	Nasal	217
	Breathy	208
	Hum-like	160
	Hiccup-like	78
	Grunt-like	17
	Speech-laugh	11
	Silencious	86

Table 2: Occurrences of the laughter sublabels

The average duration of a laughter episode is 3.6s (standard deviation: 5.5s). A histogram of the laughters durations and their cumulative distribution function is presented in Figure 3. 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 laughters duration and, above all, are the most striking ones. The longest giggle in the analyzed database lasts 82s.

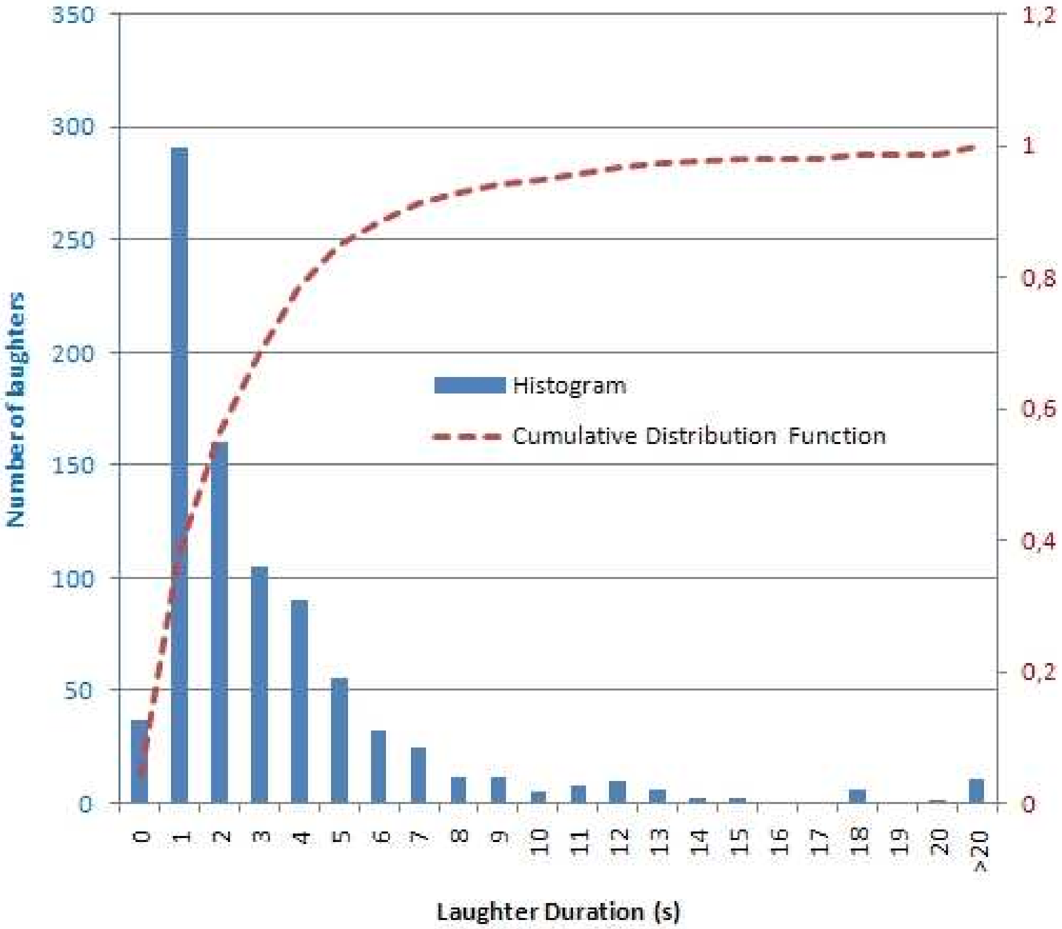

Figure 3: Histogram and cumulative distribution function of the laughters durations

To summarize, a database containing a broad variety of laughter kinds has been recorded. The presence of facial motion data in addition to the acoustic signal makes this database unique. It can be used for various research purposes: audio and/or visual laughter recognition or synthesis, etc. The corpus will soon be available from the eNTERFACE'09 website (http://www.infomus.org/enterface09/).

REFERENCES

[1] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, "Greta: an interactive expressive ECA system," C. Sierra, C. Castelfranchi, K. S. Decker, and J. S. Sichman, Eds. IFAAMAS, 2009, pp. 1399–1400.
[2] J. Wilting, E. Krahmer, and M. Swerts, "Real vs. acted emotional speech," in Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 ICSLP), Pittsburgh, USA, September 2009, pp. 805–808.
[3] K. P. Truong and D. A. van Leeuwen, "Automatic discrimination between laughter and speech," Speech Communication, vol. 49, pp. 144–158, 2007.
[4] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in Proceedings of Interspeech 2007, Antwerp, Belgium,August 2007, pp. 2973–2976.
[5] L. Kennedy and D. Ellis, "Laughter detection in meetings," in NIST
ICASSP 2004 Meeting Recognition Workshop, Montreal, May 2004.
[6] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong-Kong, April 2003.
[7] J. Wagner, E. André, and F. Jung, "Smart sensor integration: A framework for multimodal emotion recognition in real-time," in Affective Computing and Intelligent Interaction (ACII 2009), 2009.
[8] Zign Creations, "Zign track - the affordable facial motion capture solution," http://www.zigncreations.com/zigntrack.html, Consulted on October 20, 2009.
[9] Natural Point, Inc., "Optitrack - optical motion tracking solutions," http://www.naturalpoint.com/optitrack/, Consulted on October 20, 2009.
[10] J. Trouvain, "Segmenting phonetic units in laughter," in Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain, August 2003, pp. 2793–2796.