Towards a virtual agent using similarity-based laughter production

Jérôme Urbain[1], Stéphane Dupont[1], Thierry Dutoit[1], Radoslaw Niewiadomski[2], Catherine Pelachaud[2]

[1]: TCTS Lab, Faculté Polytechnique de Mons, 31 Boulevard Dolez, 7000 Mons, Belgium

[2]： CNRS - LTCI UMR 5141, Institut TELECOM - TELECOM ParisTech, 46 rue Barrault, 75013 Paris, France

In this abstract we present a collaborative project on creating a laughing machine. The machine can automatically detect laughter. After clustering it, the machine finds the closest similar laughter that is then synthesized acoustically and visually by a virtual agent. Below we present the various components involved in our project.

Facial expressions of emotions are often described statically at their apex. Lately several researchers (Keltner, 1995) have shown through analysis of video corpora that emotions are expressed through a sequence of micro-behaviours. These micro-behaviours correspond to signals spread over the whole body (face, head, gaze, gesture, etc). All these signals do not have to occur simultaneously. Some of them occur more often at the beginning of expressions, some others at the end, some occur in a sequence, while others are synchronized (Keltner, 1995).

We have developed a language to represent expressions of emotions as temporal sequences of multimodal signals. The expression of an emotional state is defined by a *behaviour set* – a set of signals through which the emotion is displayed - and a *constraints set* that defines relations between the signals in the behaviour set. These two sets are defined from the literature and from the annotation of a video corpus.

The single signals are described in the repository files while the behaviour sets are described in a central database of behaviors called *lexicon*. It is an XML-based file that specifies the mapping between the communicative intentions of the agent and behaviour sets. The relations between signals (i.e. constraints sets) are also described in XML-like format.

From an emotion label (e.g. anger or embarrassment) our system generates multimodal expressions of the emotional state i.e. the animation of a sequence of signals over different modalities. It does it by choosing a coherent subset of signals from the behaviour set, their durations, and order of display.

The algorithm can be seen as part of **Behaviour Planner** layer of the SAIBA architecture (Vilhjálmsson et al, 2007). The emotional state of the agent is one of its communicative intentions that is described in FML (or FML-APML) language. Our model translates it to a set of behaviours described in BML.

In the near future we plan to adapt this multimodal expression representation to laugh. Specific attention will be paid to the synchronization between multimodal signals and acoustic segments.

In particular, this emotional behavior language will be used in the AVLaughterCycle project during the eNTERFACE09 1-month workshop in Genova this summer. 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. The hope is that the initially forced laughter of the user will progressively turn into spontaneous laughter.

The project will start from AudioCycle (Dupont, 2009), a pre-existing audio version of this idea developed in the context of the NUMEDIART Belgian R&D program (www.numediart.org), in which a similarity-based search engine has been designed. This system searches in a corpus of audio loops the ones that are similar to some input loop according to user-defined criteria (related to rhythm, timber, and harmony).

In AudioCycle, audio loops are visually organized as follows: samples are clustered using K-means, a reference node is centered in the space and each cluster receives a region of space around it. Each loop is then placed at a radial distance inversely proportional to its similarity with the reference node, while the angular distance is related to the similarity with the cluster centroid. AudioCycle allows the user to weight the characteristics (e.g. focusing on timbre only), to define which sets of characteristics are involved in the radial and angular distances (e.g. timbre for the radial distance, harmony for the angular one), to choose the number of clusters, to define a new reference loop around which the database must be organized, etc. Open Scene Graph (OpenSceneGraph, 2009) has been used for the visual rendering.

The user can activate any loop to hear it. The sample is then loop played and its waveform displayed. To ease the localization of regions of interest when several loops are playing together, the sounds are positioned in space in relation with the position of their visual representation on the screen. This has been implemented thanks to OpenAl (OpenAl, 2009). Loops playing simultaneously are synchronized in order to allow for coherent artistic performances.

Work is in progress to adapt this software for laughter corpus browsing. New features will be extracted. Indeed, harmony makes little sense when considering laughter and the way we computed descriptors to characterize the rhythm, based on Beats localization, is inappropriate. However, timbre characterization is an interesting property when comparing laughters. Features specific to laughter properties will be included, for example to characterize the ratio of voiced versus unvoiced segments, the "phonemes" used, the periodicity (similar to rhythm), the noisy patterns exhibited by nasal or whisper-like laughters, etc.

Nevertheless, the system has already been tested as it is on laughter, for first qualitative results. Laughters from the ICSI Meeting corpus (Janin et al, 2003), using the segmentation made by Truong (Truong and van Leeuwen, 2007) constituted the database. Timbre was given a much higher weight than harmony and rhythm was not considered. Interesting trends have been observed. First, even if the clustering is not perfect, tendencies are appearing: whisper-like occurrences, laughters that could be qualified as "retained", more melodious laughters or occurrences with a lot of background noise are roughly gathered in different clusters. Second, as expected when computing similarities based on timbre, occurrences from a single person are generally close to each other. Finally, although laughters cannot be looped as music loops, activating many laughter samples from different regions of the scene gave the impression to be surrounded by a laughing audience. These primary results are promising, since the device is not optimized for laughter at all. Apart from improving the extracted features, upcoming work will also be interested in how to efficiently "synchronize" different laughters and "loop" utterances to avoid the repetitive pattern generated for the moment and build a more realistic laughing audience.

The main challenges of the eNTERFACE09 AVLaughterCycle project will therefore be:

- To be able to extend the similarity-based search to AV laughter
- To find out how to appropriately answer to an incoming laughter, the best answering laughter not necessarily being the most similar one.
- To be able to apply the output laughter taken from the corpus to an embodied conversational agent (ECA), using the emotional behavior descriptions.

References

- S. Dupont, N. D'Alessandro, T. Dubuisson, C. Frisson, R. Sebbe and J. Urbain. Audio Cycle. To appear in : QPSR of the numediart research program, T. Dutoit and B. Macq, Editors, volume 2, number 4, 2009.
- A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters. The ICSI Meeting Corpus. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong-Kong, April 2003.
- D. Keltner. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. In P. Ekman & E. L. Rosenberg (eds.), What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Series in affective science, New York: Oxford University Press, pages 133-160, 1995.
- Openal. http://www.openal.org/. Consulted on January 29, 2009.
- Open Scene Graph (osg). http://www.openscenegraph.org/. Consulted on January 29, 2009.
- K. P. Truong and D. A. van Leeuwen. Automatic discrimination between laughter and speech. Speech Communication, 49:144-158, 2007.
- H.H. Vilhjálmsson, N. Cantelmo, J. Cassell, N.E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A.N. Marshall, C. Pelachaud, Z. Ruttkay, K.R. Thórisson, H. van Welbergen, and R.J. vander Werf. The Behavior Markup Language: Recent developments and challenges. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, editors, Proceedings of 7th International Conference on Intelligent Virtual Agents, Paris, France, volume 4722 of Lecture Notes in Computer Science, Springer, pages 99-111, 2007.