SCE in HMI
Social Communicative Events in Human Machine Interactions

Hüseyin Çakmak and Kevin El Haddad

February 24, 2016

Enhancing the human-machine interaction by adding emotions to the machine's way of expression, is one of the main topics of current research. This would improve the interaction relying on the assumption that more human-like the machine behavior is, more comfortable the interaction with it will be. This project proposes to create an environment aware emotional avatar. This avatar will be placed in an experimental framework which is described as follows: Participants will be interacting with each other in a limited room, in which the avatar will be present as a reactive spectator. The avatar should be able to recognize when different scenario events (specified scenarios described later) occur and react in an affective way with respect to each occurrence. It will react through different affective expressions and speech sentences. The impact that, adding an emotional aspect to the machine's expression, can have on the degree of interest of the participants towards the machine, will be studied.

1 Project Objectives

General objectives of this project are listed below :

- Exploring the use of Deep Neural Networks (DNN) for audiovisual synthesis of Social Communicative Events (SCE), including laughter, amusement, surprise and disgust.
- Exploring the use Recurrent Neural Networks (RNN) for audiovisual recognition of SCE.
- Applying the synthesis and recognition results on a real time 3D agent with WOZ-like controls.
- Building databases for the purpose of the previous objectives.

2 Background information

The work that will be undertaken during this workshop will be based on our previous work related to the subject. Indeed, laughter synthesis has already been investigated by the team leaders. Our work audio and audiovisual isolated laughter synthesis can be found in [1, 2]. Work on adding amusement to speech by inserting laugh bursts in it as well as adding a smiling dimension to the speech was also made [3, 4, 5, 6].

The work of this project will also be based on previously collected and available multimodal databases containing SCE such as [7, 8, 9, 10, 11].

Previous work on multimodal affect recognition can be found giving good results with respect to this task [12, 13]. These works among, others, will inspire ours on the affect recognition task.

3 Detailed technical description

3.1 Technical description

The systems that will be developed in this project for the purpose of audiovisual SCE synthesis and multimodal SCE recognition will all be based on data. Therefore a data collection step will come first. In this stage, we will gather data relevant to our purposes from already available databases. The eNTERFACE workshop, due to the diversity and amount of its participants, is an opportunity to collect new databases that might be useful for current and further research. Therefore, acted and naturally expressed data from actors and participants will also be recorded in order to increase our database and further improve our systems. The current data modalities of interest are (but not limited to):

- **Audio**: The audio signal will be resampled to the most suited frequency (to be decided after the database is gathered).
- **Video**: Captured with high and/or low quality cameras (depending on the database used).
- **Motion capture**: More precisely facial expression data captured with the Optitrac system (marker based) and the Kinect system (markerless based).

As mentioned earlier, the finally obtained database will be used to develop an SCE multimodal recognition system as well as an SCE synthesis system. During this workshop we will investigate the use of the deep learning technologies as well as other machine learning systems in order to achieve our goals. The considered, but not limited to systems, are given in Table 3.1 along with the task they will be used for.

System	**Examples**	**Task**
Deep Learning systems	LSTM, DBM, DBN, multilayer ELM	Synthesis and recognition
Regression analysis	LDA, Logistic regression	Recognition
Others	SVM, ELM	Recognition

3.2 Synthesis

The Audio and visual synthesis systems will be developped separately for the different SCE. Concerning the audio cue, the features extracted in order to train the systems will be commonly used features in synthesis such as the fundamental frequency and the Mel-Frequency Cepstral Coefficiants (MFCC). These also proved efficient for synthesizing non-verbal sounds [3, 2]. Features from the motion capture data recorded by the Optitrac system will be used to train the systems. The features will be principal components obtained from a Principal Component Analysis (PCA) applied on the coordinates of each marker given by the system. This previously proved efficient [10].

3.3 Recognition

For recognition, the audio, motion capture and video cues will be considered. Fusion at both feature and decision levels will be considered and investigated for our task. The ideal object would be to obtain a system working in real time. So, the work will be focused on obtaining the most efficient system with respect to that goal.

Ideally the synthesis and recognition systems will be used and interfaces in a Human-Agent interaction experiment. A simple rule based system would be used to decide connect the recognition results to the synthesis module. The idea would be to mimic the recognized SCE using a real time 3D agent with WOZ-like controls previously developed during eNTERFACE'15. The agent has two main roles. Firstly it is the medium used to assess and demonstrate the visual synthesis. Secondly, it will be the face of the machine in a possible human-machine interaction scenario during the project.

3.4 Resources needed

The resources needed for this project are:

- **Computers** *Each participant should bring his own computer*
- **Microphones** *provided by team leaders*
- **Optitrac System** (12 infrared cameras) *provided by team leaders*
- **Kinect One** *provided by team leaders*
- **Webcams** *provided by team leaders*
- **Room for the recordings**

4 Work plan and implementation schedule

The implementation schedule that we intend to follow is as follows :

WEEK1 : The milestones of the first week include the possible building of a new database as described in Section 3. The state of the art and training of the team members (if needed) to the chosen techniques is also among the objectives of the first week.

WEEK2 : The second week should start (if not earlier) with implementation or development of the chosen techniques for the topics that will be addressed among recognition and synthesis of different types of SCE.

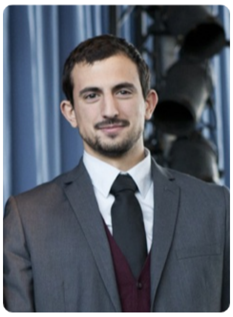
WEEK3 : The third week would be the prototyping week during which developments and implementations should be terminated.

WEEK4 : The final week would be used for polishing the codes and packaging. Preparing the demos and evaluation results for presentation.

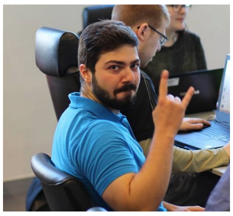
5 Benefits of the research

Research in the field of SCE processing and more specifically non-verbal vocalization and co-speech gestures analysis and synthesis remains a relatively scarce. Any advances in the field would add value to the state of the art. This project also has many common points with the European project Joker to which the advances in this project would benefit as well. Finally, advanced techniques which are in many research trends such as LSTM-DNNs are intended to be studied for the specific case of SCE recognition and synthesis. We believe that the combination of the above elements would produce novel and valuable results.

6 Profile team



Dr. Hüseyin Çakmak holds a double degree in Aeronautics from the Higher Institute of Aeronautics and Space (ISAE) and in Electrical Engineering from the Polytechnic Faculty of Mons (FPMS). In 2013, he won a FRIA grant to continue with a PhD thesis. In 2016, he finished his PhD on audiovisual laughter synthesis based on a statistical approach. His research interests are audio and visual synthesis and recognition.



Kevin El Haddad holds a Master degree in microsystems and embedded systems from the Lebanese University. He is currently PhD student at the TCTS lab. of the Polytechnic Faculty of Mons (FPMS). He works on Affect Bursts analysis and synthesis in the framework of the European project JOKER.

More information on the team leaders (research topics and publications) may be found at `http://tcts.fpms.ac.be/~laughter`.

Concerning other team members, many previous eNTERFACE participants who we already collaborated with might be interested in participating to this project. The team leaders has been participating to eNTERFACE workshops since 2012 within projects strongly related to the present proposal (virtual agents (eNTERFACE 15 - 20 participants), reactive synthesis (eNTERFACE 13&14 - 20 participants), laughing avatar (eNTERFACE 12 - 15 participants). However, we do not intend to have a team too large for this project as we would like to focus on a more specific topic and allow the team leaders to dig into implementation besides their management duties.

We are looking for any participants with good knowledge in signal processing and if possible in machine learning, preferably PhD students or higher.

References

[1] J. Urbain, H. Çakmak, and T. Dutoit, "Evaluation of HMM-based laughter synthesis," in *Acoustics Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013.

[2] H. Çakmak, J. Urbain, J. Tilmanne, and T. Dutoit, "Evaluation of HMM-based visual laughter synthesis," in *Acoustics Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014.

[3] Kevin El Haddad, Stéphane Dupont, Jérôme Urbain, and Thierry Dutoit, "Speech-laughs: An HMM-based approach for amused speech synthesis," in *Acoustics Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, in press.

[4] Kevin El Haddad, Hüseyin Cakmak, Stéphane Dupont, , and Thierry Dutoit, "Breath and repeat: An attempt at enhancing speech-laugh synthesis quality," in *European Signal Processing Conference (EUSIPCO 2015)*, Nice, France, 31 August-4 September 2015.

[5] Kevin El Haddad, Stéphane Dupont, Nicolas d'Alessandro, and Thierry Dutoit, "An HMM-based speech-smile synthesis system: An approach for amusement synthesis," in *International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, Ljubljana, Slovenia, 4-8 May 2015.

[6] Kevin El Haddad, Hüseyin Cakmak, Stéphane Dupont, and Thierry Dutoit, "An HMM Approach for Synthesizing Amused Speech with a Controllable Intensity of Smile," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Abu Dhabi, UAE, 7-10 December 2015.

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, JeannetteN. Chang, Sungbok Lee, and ShrikanthS. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[8] G Cowie McKeown, "R., curran, w., ruch, w., douglas-cowie e., ilhaire laughter database," in *Proceedings of the 4th International Workshop on EMOTION SENTIMENT & SOCIAL SIGNALS*, 2012.

[9] S. Petridis, B. Martinez, and M. Pantic, "The mahnob laughter database," *Image and Vision Computing Journal*, vol. 31, no. 2, pp. 186–202, February 2013.

[10] H. Çakmak, J. Urbain, and T. Dutoit, "Hmm-based synthesis of laughter facial expression," *Transactions on Affective Computing (TAC)*, 2015, [Submitted].

[11] L. Devillers, S. Rosset, G. Dubuisson Duplessis, M. A. Sehili, L. Bechade, A. Delaborde, C. Gossart, V. Letard, F. Yang, Y. Yemez, B. B. Turker, M. Sezgin, K. El Haddad, S. Dupont, D. Luzzati, Y. Esteve, E. Gilmartin, and N. Campbell, "Multimodal Data Collection of Human-Robot Humorous Interactions in the JOKER Project," Xi'an, China, 21-24 September 2015.

[12] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, pp. 42–55, April 2012, Issue 1.

[13] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, Jan 2009.