The First Audio/Visual Emotion Challenge and Workshop – An Introduction

Björn Schuller[1], Michel Valstar[2], Roddy Cowie[3], and Maja Pantic[2,4]

[1] Technische Universität München,
Institute for Human-Machine Communication, Munich, Germany
schuller@tum.de
[2] Imperial College London, Intelligent Behaviour Understanding Group, London, UK
michel.valstar@imperial.ac.uk
[3] Queen's University, School of Psychology, Belfast, BT7 1NN, UK
r.cowie@qub.ac.uk
[4] Twente University, EEMCS, Twente, The Netherlands
m.pantic@imperial.ac.uk

Abstract. The *Audio/Visual Emotion Challenge and Workshop (*http://sspnet.eu/avec2011*)* is the first competition event aimed at comparison of automatic audio, visual, *and* audiovisual emotion analysis. The goal of the challenge is to provide a common benchmark test set for individual multimodal information processing and to bring together the audio and video emotion recognition communities, to compare the relative merits of the two approaches to emotion recognition under well-defined and strictly comparable conditions, and establish to what extent fusion of the approaches is possible and beneficial. A second motivation is the need to advance emotion recognition systems to be able to deal with naturalistic behavior in large volumes of un-segmented, non-prototypical and non-preselected data as this is exactly the type of data that real systems have to face in the real world. Three emotion detection sub-challenges were addressed: emotion detection from audio, from video, or from audiovisual information. As benchmarking database the SEMAINE database of naturalistic dialogues was used. Emotion needed to be recognized in terms of positive/negative valence, and high and low activation (arousal), expectancy, and power.

In total, 41 research teams registered for the challenge. The data turned out to be challenging indeed: The dataset consists of over 4 hours of audio and video recordings, 3,838 words uttered by the subject of interest, and over 1.3 million video frames in total, making it not only a challenge to detect more complex affective states, but also to deal with the sheer amount of data.

Besides participation in the Challenge, papers were invited addressing in particular the differences between audio and video processing of emotive data, and the issues concerning combined audio-visual emotion recognition.

We would like to particularly thank our sponsors – the Social Signal Processing Network (SSPNet), and the HUMAINE Association, all 22 members of the Technical Program Committee for their timely and insightful reviews of the submissions: Anton Batliner, Felix Burkhardt, Rama Chellappa, Mohamed Chetouani, Fernando De la Torre, Laurence Devillers, Julien Epps, Raul Fernandez, Hatice Gunes, Julia Hirschberg, Aleix Martinez, Marc Mehu, Marcello Mortillaro, Matti Pietikäinen, Ioannis Pitas, Peter Robinson, Stefan Steidl, Jianhua Tao, Mohan Trivedi, Matthew Turk, Alessandro Vinciarelli, and Stefanos Zafeiriou, and of course all participants.