SEMAINE

THE SENSITIVE AGENT PROJECT

D3b

Human affect analyser

Date: 22 December 2009

Dissemination level: Public

ICT project contract no.	211486
Project title	**SEMAINE** **Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression**
Contractual date of delivery	*31 December 2009*
Actual date of delivery	*22 December 2009*
Deliverable number	D3b
Deliverable title	Human affect analyser
Type	Demonstrator
Number of pages	9
WP contributing to the deliverable	WP 3
Responsible for task	Maja Pantic (m.pantic@imperial.co.uk)
Author(s)	Florian Eyben, Hatice Gunes, Maja Pantic, Björn Schuller, Martin Wöllmer
EC Project Officer	Philippe Gelin

Table of Contents

1 Executive Summary

Sensitive Artificial Listeners (SAL) are virtual dialogue partners who, despite their very limited verbal understanding, intend to engage the user in a conversation by paying attention to the user's emotions and non-verbal expressions. The SAL characters have their own emotionally defined personality, and attempt to drag the user towards their dominant emotion, through a combination of verbal and non-verbal expression.

The SEMAINE system 2.0 is the first public demonstrator of the fully operational autonomous SAL system based on audiovisual analysis and synthesis. The present report is part of a group of reports describing various aspects of the SEMAINE system 2.0. The full list of reports is available from http://semaine.opendfki.de/wiki/SEMAINE-2.0.

This report describes the current state of the human affect analysers integrated into the SEMAINE system 2.0, namely, valence and arousal analyser from facial feature points tracked, analysis of valence, arousal, power, anticipation/expectation and intensity dimensions from the audio features extracted, and the investigation of the level at which multiple visual cues (e.g., head motion and facial expressions) and/or audio and visual modalities (e.g., for laughter detection) should be fused.

2 System description

2.1 Automatic Recognition of Arousal and Valence from Tracked Facial Feature Points

The data for training the automatic recognizer were obtained from the SAL database. After the audiovisual SAL data have been automatically segmented and appropriate annotations from multiple coders have been obtained, Support Vector Machines for Regression (SVR) were trained for recognition of arousal and valence in every video frame, based on the tracked positions of the facial feature points in every frame. For our experiments, we used both polynomial (SVR-P) and radial-basis function (SVR-RBF) kernels. For recognition of valence, best results with respect to the mean squared error (MSE), for both polynomial and RBF kernels was MSE=0.054. For recognition of arousal, the polynomial kernel provided slightly better results with MSE=0.088 compared to that of the RBF kernel (MSE=0.090). As the difference between the MSE of each kernel is very small (0.002), we could not draw significant conclusions on which function performs better.

We therefore investigated other metrics to evaluate the performance of dimensional approaches to automatic and continuous affect recognition, namely, (i) the correlation (COR) of the ground truth (with the estimated values) and (ii) the agreement (AGR) which is the percentage where the estimated value agrees with the ground truth value.

When we analysed the performance of different SVR kernels in terms of the above mentioned measures other than MSE, we observed that SVR-RBF provided COR=0.159 and AGR=0.549 whereas SVR-P provided COR=0.025 and AGR=0.490 for recognition of arousal. For recognition of valance, SVR-RBF provided COR=0.091 and AGR=0.618 whereas SVR-P provided COR=0.002 and AGR=0.678. Hence, for the Semaine system, we opted for the use of SVR-RBF for automatic recognition of arousal and valence. These results provide important suggestions on how automatic dimensional affect recognizers should be trained and evaluated using the newly acquired and annotated SEMAINE data.

The automatic arousal and valence recognizer is implemented in the ArousalValenceAnalyser module of the SEMAINE framework. Information about the arousal and valence is sent to the data.state.user.emma topic as an EMMA message. The frequency of this message can be set (in terms of number of video frames) by modifying the config file. The EMMA message is only sent if a face and its feature points have been detected continuously for the predefined number of frames.

2.2 Automatic Recognition of Valence, Arousal, Power, Anticipation/Expectation and Intensity from Extracted Audio Features

Recognition of emotional dimensions from audio features is a fairly novel field of research. Only few corpora exist where emotion dimensions are labelled time-continuously. Besides the SEMAINE database, the only corpus which is usable for the SEMAINE system is the SAL corpus and to some extent the Vera-Am-Mittag corpus. The latter does not contain fully continuous annotations generated with FEELTrace. Instead, self-assessment manikins were used to produce discrete valence and arousal ratings for each rater. These ratings were averaged over all raters, thus producing pseudo continuous values, which are sub-optimal if recognition of a true continuum is desired. In addition, these are only available per turn, while sub-turn entities seem more suited (Batliner et al., 2009). We thus conducted extensive experiments (Wöllmer et al, 2008; Eyben et al., 2009; Schuller et al., 2009; Wöllmer et al., 2009) with the dimensions arousal and valence on the SAL data because it

best matches to the scope of the SEMAINE system. Using the same features and classifiers, models for the five dimensions Valence, Arousal, Power, Anticipation, and Intensity annotated in the new SEMAINE data were trained and included in the SEMAINE 2.0 release.

The recognition techniques used are similar to those used for recognition of emotion categories: functionals are applied to contours of low-level descriptors, thus mapping these contours of variable length to a vector of fixed dimensionality. Experiments were conducted on the SAL data which contains two dimensions: arousal and valence. In our experiments Long Short-Term Recurrent Neural Networks (LSTM-RNN) showed good performance in (Wöllmer et al, 2008) for turn-based emotion recognition and in (Eyben et al., 2009b) for frame-based emotion recognition, which has not been attempted before. Yet, the Long Short-Term memory networks have not been implemented into the live system. Therefore, the SEMAINE 2.0 system uses Support Vector Regression models, which have also shown good performance in our studies (Eyben et al., 2009; Schuller et al., 2009). The Support Vector Regression uses a polynomial kernel function of degree 1. We also investigated RBF kernels, however our experiments in line with our long standing experience in the field has shown that the best performing SVR kernel for emotion recognition from audio is the polynomial kernel. Moreover, the polynomial kernel is faster to evaluate than the RBF kernel, which is beneficial for fast processing and responsiveness in the SEMAINE system.

The SEMAINE system must be able to respond to emotional states very fast. Conventional turn-based emotion recognition approaches output one result at the end of the turn, which causes a non acceptable delay in some cases. This is solved by incremental output of the current best guess. Every second an estimate of the current emotional state is sent. This estimate is obtained by applying functionals to low-level descriptor contours in a window from the beginning of the turn to the current position. The resulting feature vector is then classified using the turn-based models. In addition, the SEMAINE system naturally has to process all speech with respect to emotional content and not only prototypical turns, as usually found in the literature. In this respect studies were carried out to show the impact of processing all input and strategies to best cope with this situation (Schuller et al., 2009b; Steidl et al., 2009). These studies indicate a clear downgrade of recognition performance for more naturalistic und non-prototypical data.

Next to acoustic features, the affect recognition module in the SEMAINE 2.0 system uses linguistic cues to improve the estimates for arousal and valence. This has improved recognition especially for valence (Eyben et al., 2009b) and was found to be surprisingly invariant to colouring of the speech throughout speech recognition (Schuller et al., 2009c). Using automatic feature selection methods, 21 keywords out of a list of 56 keywords were found to be correlated to arousal, while 40 out of the 56 are correlated to valence. Thus one binary bag-of-words vector for each of the two emotion dimensions is constructed: for all keywords which appear in the current turn, the corresponding elements in the linguistic feature vector are set to 1, all other elements remain 0. Since the keyword spotter output is also incremental, the component which builds the bag-of-words vector stores the most current keyword spotter output and thus achieves good synchronisation between acoustic and linguistic features at minimal cost.

The affect recognisers are currently part of the SEMAINE component TumFeatureExtrator. The emotion estimates are sent every two seconds and at the end of a user turn to the topic state.user.emma as Emma XML messages containing EmotionML.

3 License and availability

The automatic arousal and valence recognizer from visual cues is available as freeware as a binary executable functioning as part of the overall SEMAINE system 2.0. It is also available separately, as an installable executable together with all other components that rely on visual information processing and analysis.

Affect recognition modules are available under the terms of the GPL within the SEMAINE system (included in SEMAINE download package; linux and windows versions available). The classification uses the open-source third-party library LibSVM for support vector classification, which is distributed under a BSD-style license. The affect recognition module is also available as a standalone open-source package (openEAR) to the emotion research community.

References

(Batliner et al., 2009) A. Batliner, D. Seppi, S. Steidl, B. Schuller: “Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach”, in Advances in Human Computer Interaction, Special Issue on "Emotion-Aware Natural Interaction" (to appear), 2009.

(Eyben et al., 2009) F. Eyben, M. Wöllmer, B. Schuller, “openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit”, Proc. of 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII 2009), Amsterdam, The Netherlands, IEEE, pp. 576-581, 2009.

(Eyben et al., 2009b) F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, R. Cowie, “On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues”, in Journal on Multimodal User Interfaces, Special Issue on Real-time Affect Analysis and Interpretation: Closing the Loop in Virtual Agents (to appear), Springer, 2009.

(Petridis et al., 2009) S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic, "Static vs. Dynamic Modeling of Human Nonverbal Behavior from Multiple Cues and Modalities", Proc. of ACM Int’l Conf. on Multimodal Interfaces, pp. 23-30, 2009.

(Schuller et al., 2009) B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll and A. Wendemuth: Acoustic Emotion Recognition: A Benchmark Comparison of Performances, in: Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU), Merano, Italy, IEEE, 2009.

(Schuller et al., 2009b) B. Schuller, S. Steidl, A. Batliner, “The Interspeech 2009 Emotion Challenge”, Proc. Interspeech, ISCA, Brighton, UK, ISCA, pp. 312-315, 2009.

(Schuller et al., 2009c) B. Schuller, A. Batliner, S. Steidl, D. Seppi, “Emotion Recognition from Speech: Putting ASR in the Loop”, Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009), Taipei, Taiwan, IEEE, pp. 4585-4588, 2009.

(Steidl et al., 2009) S. Steidl, B. Schuller, D. Seppi, A. Batliner, “The Hinterland of Emotions: Facing the Open-Microphone Challenge”, Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII 2009), Amsterdam, The Netherlands, IEEE, pp. 690-697, 2009.

(Wöllmer et al, 2008) M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, “Abandoning Emotion Classes - Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies”, in Proc. of Interspeech, Brisbane, Australia, ISCA, pp. 597-600, 2008.

(Wöllmer et al., 2009) M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, R. Cowie: Data-driven Clustering in Emotional Space for Affect Recognition Using Discriminatively Trained LSTM Networks”, in Proc. of Interspeech, Brighton, UK, ISCA, pp. 1595-1598, 2009.

(Wöllmer et al., 2009b) M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, G. Rigoll, "A Multidimensional Dynamic Time Warping Algorithm for Efficient Multimodal Fusion of Asynchronous Data Streams", Neurocomputing, Elsevier, Vol. 73:1-3, pp. 366-380, 2009.