SEMAINE

THE SENSITIVE AGENT PROJECT

# D3c

# User-profiled human behaviour interpreter


**Date: 24 September 2010**

**Dissemination level: Public**

| | |
|---|---|
| **ICT project contract no.** | 211486 |
| **Project title** | **SEMAINE**<br>**Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression** |
| **Contractual date of delivery** | *31 August 2010* |
| **Actual date of delivery** | **24 September 2010** |
| **Deliverable number** | **D3c** |
| **Deliverable title** | **User-profiled human behaviour interpreter** |
| **Type** | Demonstrator |
| **Number of pages** | 24 |
| **WP contributing to the deliverable** | WP 3 |
| **Responsible for task** | Maja Pantic (m.pantic@imperial.ac.uk) |
| **Author(s)** | Florian Eyben, Hatice Gunes, Maja Pantic, Marc Schroeder, Björn Schuller, Michel F. Valstar, 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.

This report is part of the series of reports describing the implementation of SAL in system SEMAINE-3.0. The software described, and the full set of reports, can be downloaded from http://semaine.opendfki.de/wiki/SEMAINE-3.0/.

This report describes the functionality of the components in the SAL system, and an assessment of the quality on the technical and component level.

# 2 Functionality of the components

This section describes the functionality of the components in the SAL system. The possibilities to configure and reuse the components as parts of a research toolbox will be published as deliverable D7e in December 2010.

## 2.1 Video components

This section describes the functionality of the video components in the SAL system.

### 2.1.1 The VideoFeatureExtractor component

This component provides functionality for all tasks that need to work on the acquired image data itself. Where possible, tasks have been split into the generation of a low-dimensional signal from this image data, which can then be further analysed in another module. Good examples of this are the detection of nods and shakes and of head tilts, where information about the head motion and head pose are computed in the VideoFeatureExtractor module and sent to a specific topic so that the nodshakeAnalyser and headPoseAnalyser components can detect these gestures.

The VideoFeatureExtractor component sends four signals to the framework: the detected face location, 2D head motion estimation, head pose estimation, and appearance-based Action Unit (AU) detection. The face location and 2D head motion estimation have been explained in detail in SEMAINE-2.0.

The head pose estimation uses the face detection result combined with eye detection within the detected face region to estimate the 3-Dimensional head position relative to the camera as well as the head roll. The location of the detected eyes can be used to estimate the head roll \alpha as follows:

\alpha = \atan (\frac{y_l-y_r}{x_l-x_r})

where x_l and x_r are the horizontal position of respectively the left and right eye, and y_l and y_r are their vertical positions. The head roll is send to the topic "semaine.data.analysis.features.video.headpose", as a feature message.

The appearance-based Action Unit detection uses Uniform Local Binary Pattern histograms, that are computed from each cell of a 10 x 8 grid placed over the detected face. The resulting feature vector is given to a series Support Vector Machine (SVM) classifiers, one for every trained AU. Currently the system has trained SVMs for the AUs AU1 (raised inner eye-brow), AU2 (raised outer eye-brow), AU4 (lowered eye-brows), AU12 (smile), and AU25 (lips parted).

### 2.1.2 The facePresenceAnalyser component

The facePresenceAnalyser component takes information from the face detector, and is an integrated component of the VideoFeatureExtractor module. It does not listen to any topic, instead it is directly uses the output of the face detector (fail/success) as its input. It analyses a number of frames that correspond to approximately one second in time, and if a face is detected successfully in any frame during that period, it will send a message that the face is present to the topic "semaine.data.state.user.emma.nonverbal.face".

The confidence that corresponds to this message is the number of frames in which a face is detected divided by the number of frames in the analysed period.

### 2.1.3 The nodshakeAnalyser component

The nodshakeAnalyser component analyses the head gestures (i.e., head nods and head shakes) of the user based on 2D global head motion estimation and trained nod and shake models (explained in detail in SEMAINE-2.0). Information about the head nod and head shake is sent to the semaine.data.state.user.emma.nonverbal.head topic as an EMMA message with a frequency of 0.4 seconds (i.e., every 20 frames). The EMMA message is only sent if a nod/shake event has been detected.

The results of the head gesture analysis (head motion features, detected nod/shake) are then utilised to obtain five dimensional emotion prediction (in arousal, expectation, intensity, power and valence dimensions). Support Vector Regression (epsilon SVR with an RBF kernel) is used for dimensional emotion prediction. Please see (Gunes & Pantic, 2010b) for details.

The current version of the system uses multiple predictors (15 in total, representing 3 separate raters and 5 dimensions) trained by using the ratings from each human rater and each dimension, separately (e.g., arousal-predictor 1 is trained using ground truth values provided by rater 1 for arousal dimension; arousal-predictor 2 is trained using ground truth values provided by rater 2 for arousal dimension, etc.). More specifically, at any given time, 3 EMMA messages, containing prediction values for five dimensions are sent.

Information about the emotion is sent to the semaine.data.state.user.emma.emotion.head topic as an EMMA message with a frequency of 0.4 seconds (i.e., every 20 frames). The EMMA message is only sent if a face has been detected continuously for the predefined number of frames.

### 2.1.4 The headPoseAnalyser component

The headPoseAnalyser component analyses information about the head gestures tilt-left, and tilt-right. The headPoseAnalyser component reads from this topic, and uses the information about the head roll to detect whether the head is upright, tilted to the left, or tilted to the right. To increase robustness of the head tilt recognition, be able to attach a confidence to the gesture recognition, and to reduce the frequency at which meaningful messages are sent by this component, it collects head roll data for n frames, which is set to reflect approximately one second of video. It then calculates the confidence for a head tilt by averaging the roll over the n frames in the temporal window, where the confidence per frame is set to 1 if the roll is greater than 45 degrees, and to 0 if the roll is in the wrong direction (depending on whether we're calculating confidence for a left or right head-tilt). A head tilt is considered to occur if the confidence is greater than 0.2.

Information about these two head gestures is sent as an EMMA message to the topic "semaine.data.state.user.emma.nonverbal.head". Messages are sent approximately once per second.

### 2.1.5 The actionUnitAnalyser component

The actionUnitAnalyser component is a postprocessing module that increases the accuracy of the LBP-based Action Unit (AU) detector that is integrated in the VideoFeatureExtractor module. It listens to the "semaine.data.analysis.features.video.lbpfacs" topic, where the VideoFeatureExtractor sends the confidences of all AUs (see D2B, section ??).

To increase robustness of the AU detection, to be able to refine the confidence measure, and to reduce the frequency at which meaningful messages are sent by this component the actionUnitAnalyser analyses the incoming AU messages over a number of frames, set to be equivalent to approximately one second. The confidence of each AU is updated to be the average confidence of incoming messages over the time window. An AU is determined to be active by the analyser if its confidence is greater than 0.5.

Information about the detected AUs is sent as an EMMA message to the topic "semaine.data.state.user.emma.nonverbal.face". Messages are sent approximately once per second.

### 2.1.6 The EmotionFusion component

The EmotionFusion component has been implemented as a very generic component that can be used to fuse a very broad range of emotion messages (including all types of emotion annotations that occur in SEMAINE).

This component listens to all subtopics of semaine.data.state.user.emma.emotion.*, and computes consolidated statements about user emotion by computing weighted averages of the values, where the *confidence* is used as the weight. This works for all dimensions, appraisals, action-tendencies, categories, and for intensity. For a given emotion category, the resulting confidence is the average of the incoming confidences; for a given emotion dimension instead, the resulting confidence is one minus the standard deviation of the value across data points (variance being computed using confidence as the weight).

Two config settings (in the java config file used to start the system) can be used to influence the behaviour of the EmotionFusion:

- semaine.emotion-fusion.max-delay -- specifies the maximum waiting time, in milliseconds, before the fused data is sent (default: 500 ms).
- semaine.emotion-fusion.num-raters -- can be used to indicate a number of "expected" data points per emotion annotation; as soon as this number is reached, the fused data is sent (default: 999, i.e. this will not be reached unless explicitly set by a user).

## 2.2 Audio components

This section describes the functionality of the audio components in the SAL system.

### 2.2.1 The acousticFeatureExtractor component

The acousticFeatureExtractor component extracts acoustic features used for recognition of the user's affective state (5 continuous dimensions and 3 levels of interest). This component is not a standalone component in the SEMAINE system, it is part of the TumAudioFeatureExtractor component, which also extracts low-level prosodic features, features for the keyword spotter, and features for the voice activity detector.

For affect recognition, three different feature sets have been evaluated and all three are included for use in the live system.

The original set of features (Set A) is based on the feature set used for the baseline results of the INTERSPPECH 2010 Paralinguistic Challenge (Schuller et al., 2010a). It has been extended by 7 RASTA-PLP descriptors and 14 Mel-Frequency Bands instead of only 8 as in the challenge set (covering the same frequency range from 20-6500 Hz). There are 2,128 features in this set. The computational complexity of this set is quite high (rtf ~0.5 on a single core of a 2 GHz AMD64 machine), thus it is only suitable for the live system if the audio feature extraction is run on a powerful and fast machine.

We therefore decided to include a less computationally intense set, which also reuses features already extracted for keyword spotting and prosodic cues (see D2b report). We call this Set B. A third set with a further reduced set of features is also included, which can be used on slow machines (Set C). A description of the feature sets can be found in the tables below.

Set A (2,128 dimensional)

| Low-level descriptors (47) | Functionals (21) |
|---|---|
| Loudness | 1% and 99% percentile |
| Probability of voicing | Range of 1% and 99% percentile |
| Fundamental Frequency envelope | Relative position of maximum / minimum value |
| MFCC 0-14 | Arithmetic mean |
| RASTA PLP-CC 1-7 | Linear regression (slope, offset, linear error, quadratic error) |
| MFB 1-14 | Standard deviation, Skewness, Kurtosis |
| LSP 1-8 | Quartile 1, 2, 3 and inter-quartile ranges |
| | Time the signal is above 75%, and 90% of its range |

| Low-level descriptors (4) | Functionals (19) |
|---|---|
| Fundamental Frequency (in voiced regions) | 99% percentile |
| F0 Jitter (local) | Relative position of maximum / minimum value |
| F0 Jitter (differential) | Arithmetic mean |
| Shimmer (local) | Linear regression (slope, offset, linear error, quadratic error) |
| | Standard deviation, Skewness, Kurtosis |
| | Quartile 1, 2, 3 and inter-quartile ranges |
| | Time the signal is above 75%, and 90% of its range |

+ Segment duration in seconds

+ Number of voiced segments (segments with voicing probability over a threshold)

Set B (1,882 dimensional):

| Low-level descriptors (47) | Functionals (20) |
|---|---|
| Intensity, Loudness, RMS and Log energy<br>Probability of voicing<br>Fundamental Frequency (in voiced regions)<br>MFCC 0-12<br>RASTA PLP-CC 0-7<br>MFB 1-14<br>95% Spectral Roll-off point<br>Spectral Flux<br>Spectral Centroid<br>Spectral Entropy<br>Spectral Variance<br>Frame-Mean Crossing Rate | Maximum / minimum value<br>Relative position of maximum / minimum value<br>Range<br>Arithmetic mean<br>Linear regression (slope, offset, linear error, quadratic error)<br>Standard deviation, Skewness, Kurtosis<br>Time the signal is above 25%, 50%, 75%, and 90% of its range<br>Time the signal is below 50% of its range<br>Time the signal is rising<br>Time the signal is falling |

+ Segment duration in seconds

+ Number of voiced segments (segments with voicing probability over a threshold)

Set C (1,322 dimensional)

| Low-level descriptors (33) | Functionals (20) |
|---|---|
| Intensity, Loudness, RMS and Log energy<br>Probability of voicing<br>Fundamental Frequency (in voiced regions)<br>MFCC 0-12<br>RASTA PLP-CC 0-7<br>95% Spectral Roll-off point<br>Spectral Flux<br>Spectral Centroid<br>Spectral Entropy<br>Spectral Variance<br>Frame-Mean Crossing Rate | Maximum / minimum value<br>Relative position of maximum / minimum value<br>Range<br>Arithmetic mean<br>Linear regression (slope, offset, linear error, quadratic error)<br>Standard deviation, Skewness, Kurtosis<br>Time the signal is above 25%, 50%, 75%, and 90% of its range<br>Time the signal is below 50% of its range<br>Time the signal is rising<br>Time the signal is falling |

+ Segment duration in seconds

+ Number of voiced segments (segments with voicing probability over a threshold)

### 2.2.2 The affectAnalyser component

The affectAnalyser component is able to recognise 5 dimensional affect with models trained on the SEMAINE database. The 5 dimensions thus are arousal, expectation, intensity, power, valence. The feature sets as described in the previous section have been used for all models. The component gets the features directly from the acousticFeatureExtractor component, which are both run within the openSMILE executable. This increases performance, since large feature vectors are not sent via the ActiveMQ message server.

Acoustic emotion recognition is based on a two step process. In the first step low-level audio descriptors are extracted and functionals of these descriptors are computed for short time segments in order to obtain one feature vector per segment. The values of five emotion dimensions are then predicted from each feature vector by a classifier. The user can choose between Support-Vector Machines or Long Short-Term Memory Recurrent Neural Networks (for performance comparison and evaluations, see section 3).

Segment boundaries start when the user starts speaking and end when the user stops speaking. A maximum segment length of 5 seconds is imposed, after when a new segment will be started. To obtain intermediate estimates of the user's affect, a preliminary output is generated every second. Thereby the segment is defined by a 5 second window reaching back in time from the current time. If the beginning of the user's speech is less than 5 seconds in the past, the segment is from the beginning of the user's speech to the current point in time.

Due to the high dimensionality of the feature vectors, Support-Vector regression (SVR) is chosen as the primary classifier since it is known to handle large feature spaces reliably. Alternatively LSTM-RNN (Long Short-Term Memory Recurrent Neural Networks) have been investigated as a replacement for SVR. LSTM-RNN enable modelling of the context from segment to segment. Recognition performance improvements for LSTM-RNN over SVR haven been shown in (Eyben et al., 2010a). Bidirectional LSTM networks were also investigated in (Eyben et al., 2010a), however they are unsuitable for the demonstrator system since they are non-causal. The current experiments with LSTM-RNN are preliminary and did not yet outperform the SVR approach, which has been studied more heavily. Thus we decided to not include LSTM-RNN emotion recognition in the current system.

For dimensional affect, it was decided to train SVR models and LSTM-RNN for individual raters instead of using the mean of all Feeltrace annotations. The raters 3, 5, and 6 have been chosen, since they have annotated most of the data. Recordings 3 and 5 have been left out for testing (see Section 3), while the remaining recordings (4,6,9,10,11,13,15,16,17,18,19) have been used for training. Epsilon-SVR with a polynomial kernel function of degree 1 is used. 20 SVR models are trained (4 raters *x* 5dimensions) and 3 LSTM networks (for the 3 coders) are trained with 5 outputs each (one for each dimension).

The result of the predictors is sent to the topic “semaine.data.state.user.emotion.voice”.

### 2.2.3 The interestDetector component

The interestDetector component detects the user's level of interest. Three cases are distinguished: bored, neutral, interested.

Support Vector Machine (SVM) models with polynomial kernel of degree 1 have been trained on the TUM AVIC corpus, described in (Schuller et al., 2009a). All data of of this corpus has been used for training the models in the demonstrator system. To reduce computational overhead in the demonstrator system, the same feature sets as for the 5 dimensional affect recognition are used (see Section 2.1.1).

The result of the interest detection is sent to the topic "semaine.data.state.user.emotion.voice". The result message contains the confidence for each of the three levels of interest.

# 3 Quality assessment

This section describes an assessment of the quality on the technical and component level. An assessment of the psychological quality of interactions with the overall system will be published separately, as deliverable report D6d, in December 2010.

## *3.1* Video components

### 3.1.1 The nodshakeAnalyser component

In the previous version of the Semaine system, the models were trained using data from SAL and SEMAINE databases. In the current version, the models were trained using data from the SEMAINE database *only* (152 head nod and 103 head shake clips of variable length). As both data used for training and the models trained are different, the comparison between the two versions of the system is not straightforward. Qualitatively it is possible to state that the current models are better able to represent the actual SEMAINE system (e.g., in terms of camera setup, video frame rate etc.).

In order to decide how to make the final decision, evaluation has been carried out (using the aforementioned data and adopting 10-fold cross-validation) on the following issues: (i) thresholding the normalised magnitude (normalised by the height of the detected face) of the head motion (0~30), (ii) deciding on the number of states to be used within the HMM models (2~5), and (iii) whether to use maximum likelihood classification (i.e., decision is based on the model that provides the maximum likelihood) or likelihood space classification (i.e., decision is made by a classifier trained using the likelihoods outputted by all HMM models). For discrete emotion recognition (Nicolaou et al., 2010a) have shown that likelihood space classification greatly improves the recognition accuracy.

The table below presents the best results. As the table shows, the best results were obtained by thresholding head motion magnitude (threshold=15 or threshold=25), and by using either 4 or 2 states within the HMM models. To keep the model and computational complexity simpler, in the current system, we opted for likelihood space classification, setting the threshold=25, and number of states=2.

| Threshold for normalised head motion magnitude | Number of states used in HMM model | Likelihood space classification (%) | Maximum likelihood classification (%) |
|---|---|---|---|
| 15 | 4 | **92.8** | 86.5 |
| 25 | 2 | **92.2** | 84.1 |
| 0 | 3 | 91.2 | 86.9 |
| 15 | 3 | 89.4 | 83.3 |
| 0 | 2 | 88.7 | 85.0 |

Dimensional emotion prediction from head gestures has been obtained by training Support Vector Regressors (epsilon SVR with an RBF kernel). The trained models were evaluated using the Semaine database, using the sessions that have been coded by 3 raters (rater 3, rater 5 and rater 6 who coded the maximum number of sessions). Recordings 4, 6, 9, 10, 11, 13, 15, 16, 17, 18, and 19 were used for training, and recording 3 and recording 5 were used for testing.

The Mean Squarred Error (MSE) and correlation are used as metrics for evaluation. MSE measures the average of the square of the error between an estimator and the true value of the quantity being estimated. Correlation (usually refers to Pearson's correlation) indicates the strength of a linear relationship between two variables.

MSE and correlation have been calculated both for raters and the (automatic) predictors. Results for each rater (or predictor) - dimension combination are presented in the table below. Both MSE and correlation for each rater has been calculated with respect to other raters and by averaging the obtained results.

| Rater/predictor-dimension | Human rater MSE | Human rater correlation | Automatic predictor MSE | Automatic predictor correlation |
|---|---|---|---|---|
| R3-DA | 0.2273 | 0.4488 | 0.0708 | 0.2962 |
| R3-DE | 0.2672 | 0.1867 | 0.1719 | -0.0107 |
| R3-DI | 0.0978 | 0.4389 | 0.0765 | 0.2255 |
| R3-DP | 0.4022 | 0.2023 | 0.4343 | 0.0175 |
| R3-DV | 0.0929 | 0.4638 | 0.0989 | 0.1105 |
| | | | | |
| R5-DA | 0.2131 | 0.4151 | 0.1676 | 0.2379 |
| R5-DE | 0.3852 | 0.1386 | 0.1879 | -0.1126 |
| R5-DI | 0.1162 | 0.4537 | 0.1227 | 0.3482 |
| R5-DP | 0.2987 | 0.2188 | 0.2649 | 0.1843 |
| R5-DV | 0.0811 | 0.3664 | 0.0519 | 0.0140 |
| | | | | |
| R6-DA | 0.1461 | 0.3815 | 0.0345 | 0.0696 |
| R6-DE | 0.3102 | 0.1097 | 0.0849 | 0.0282 |
| R6-DI | 0.0898 | 0.3012 | 0.0262 | 0.0435 |
| R6-DP | 0.4561 | 0.1837 | 0.0663 | -0.07492 |
| R6-DV | 0.0734 | 0.4273 | 0.0441 | -0.1168 |

The table illustrates that obtaining a high correlation between various human raters, for the audio-visual SEMAINE data, is indeed challenging. To mitigate a similar problem, (Nicolaou et al., 2010b) proposed a method for achieving high inter-rater agreement and segmenting the continuous sequences into shorter clips. However, extending this method for automatic, dimensional and continuous emotion prediction remains a challenge.

Moreover, during annotation, the human raters were exposed to audio-visual SEMAINE data without explicitly being instructed to pay attention to any particular cue (e.g., head gestures) or

modality (e.g., audio or video). Therefore, it is not possible to state which cues or modalities were dominant for which ratings. In general, these issues still remain as open research questions in the field (Gunes & Pantic, 2010a).

### 3.1.2 The headPoseAnalyser component

Unfortunately the SEMAINE database has not been annotated for left and right head-tilt gestures. It is therefore not possible to evaluate the performance of the headPoseAnalyser on the SEMAINE data. We also did not find any publicly available database for which the left and right head-tilt gestures have been annotated. However, by visual inspection, the headPoseAnalyser performs well.

### 3.1.3 The actionUnitAnalyser component

The system has been trained on images taken from both the MMI Facial Expression Database and the SEMAINE corpus. It was evaluated in a subject-independent manner, with results obtained for the two corpi separately. Results are shown below in the two tables. Table 1 shows the results for data taken from the MMI Facial Expression Database, and Table 2 shows the results for the data taken from the SEMAINE corpus.

| AU | Classification rate | Recall | Precision | F1-Measure |
|---|---|---|---|---|
| 1 | 0.841 | 0.588 | 0.900 | 0.580 |
| 2 | 0.787 | 0.663 | 0.756 | 0.610 |
| 4 | 0.808 | 0.564 | 0.745 | 0.483 |
| 12 | 0.86 | 0.707 | 0.846 | 0.693 |
| 25 | 0.709 | 0.727 | 0.761 | 0.697 |

Table 1. Action Unit detection performance on data from the MMI Facial Expression Database

| AU | Classification rate | Recall | Precision | F1-Measure |
|---|---|---|---|---|
| 1 | 0.840 | 0.233 | 1.000 | 0.318 |
| 2 | 0.769 | 0.042 | 0.667 | 0.067 |
| 4 | 0.649 | 0.464 | 0.741 | 0.032 |
| 12 | 0.477 | 0.500 | 0.761 | 0.405 |
| 25 | 0.514 | 0.606 | 0.603 | 0.604 |

Table 2. Action Unit detection performance on data from the SEMAINE corpus

One of the aims of the nonverbal behaviour recognition modules was to focus on delivering reliable results, rather than returning detected gestures of which many may be false positives. In performance measure terms, this means the aim was to obtain a high precision, at the cost of recall and F1-Measure. As can be seen from the tables above, this goal was indeed achieved.

A question one could ask is how the AU detection component performs for different subjects. As stated above, the system was trained and tested in a subject independent manner. We therefore add results for each subject independently in Figure 1 below. From the figure we can see that, although the performance varies somewhat between subjects, all subject attain a fairly high precision rate, and no subject has a zero recall, meaning that for every subject we do detect AUs. Note that the figure shows the average performance over all AUs for a given subject. Subjects 2 and 3 are from the SEMAINE corpus, all others are taken from the MMI Facial Expression Database.

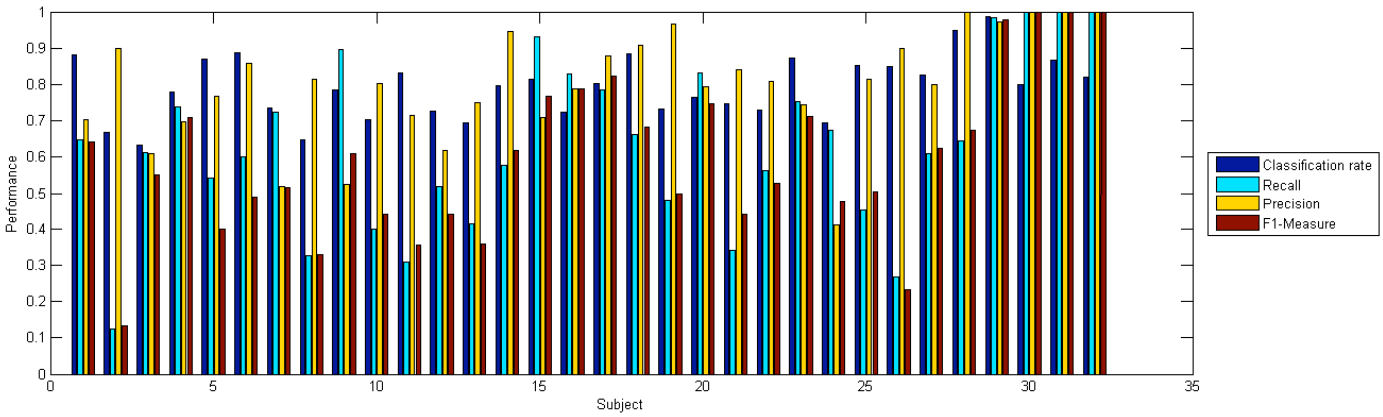


Figure 1. Average AU-detection performance per subject. Subjects 2 and 3 are taken from the SEMAINE corpus, all others from the MMI Facial Expression Database

## *3.2* Audio components

**The affectAnalyser component**

Five dimensional affect analysis from speech is evaluated using SVR with polynomial kernel function of degree 1. This has shown to give the best speaker independent results. Although higher order kernel functions give good results for speaker dependent recognition, the performance for a speaker independent test is reduced.

Here only speech turns were used for evaluation, thus the Human rater results differ from those in section 3.2.

We give results for each of the three feature sets: a) for all features in the set, b) for a feature sub-set computed via correlation-based feature selection for each rater and dimension, and c) for the feature sub-set from (b) with an appended linguistic and non-verbal feature vector (Bag-of-words (BOW) including laughter and sigh). The list of words in this vector was estimated for each dimension separately from the training set based on the mean label of the three coders.

| Rater/predictor-dimension | Human rater MSE | Human rater correlation | Automatic predictor MSE | Automatic predictor correlation |
|---|---|---|---|---|
| R3-DA | 0.1843 | 0.6145 | 0.0625 | 0.3709 |
| R3-DE | 0.1813 | 0.3578 | 0.2423 | 0.0760 |
| R3-DI | 0.0766 | 0.4942 | 0.0597 | 0.4155 |
| R3-DP | 0.3233 | 0.4942 | 0.4058 | 0.2356 |
| R3-DV | 0.0725 | 0.4942 | 0.1248 | 0.1387 |
| | | | | |
| R5-DA | 0.1920 | 0.4942 | 0.1190 | 0.0798 |
| R5-DE | 0.3700 | 0.3228 | 0.1868 | 0.0482 |
| R5-DI | 0.1226 | 0.5104 | 0.0601 | 0.4539 |
| R5-DP | 0.1341 | 0.3063 | 0.2134 | 0.0849 |
| R5-DV | 0.0585 | 0.4898 | 0.0836 | -0.1317 |
| | | | | |
| R6-DA | 0.0585 | 0.4303 | 0.0314 | 0.1661 |
| R6-DE | 0.3050 | 0.4303 | 0.0614 | 0.3169 |
| R6-DI | 0.0614 | 0.1908 | 0.0220 | 0.1106 |
| R6-DP | 0.3752 | 0.2467 | 0.0441 | 0.1704 |
| R6-DV | 0.0512 | 0.5498 | 0.0610 | 0.1597 |

Feature Set A (all) , SVM

| Rater/predictor-dimension | Human rater MSE | Human rater correlation | Automatic predictor MSE | Automatic predictor correlation |
|---|---|---|---|---|
| R3-DA | 0.1843 | 0.6145 | 0.0634 | 0.3603 |
| R3-DE | 0.1813 | 0.3578 | 0.2304 | 0.2281 |
| R3-DI | 0.0766 | 0.4942 | 0.0990 | 0.2167 |
| R3-DP | 0.3233 | 0.4942 | 0.4413 | 0.1709 |
| R3-DV | 0.0725 | 0.4942 | 0.1114 | 0.1028 |
| | | | | |
| R5-DA | 0.1920 | 0.4942 | 0.1204 | 0.0848 |
| R5-DE | 0.3700 | 0.3228 | 0.1802 | 0.0883 |
| R5-DI | 0.1226 | 0.5104 | 0.0640 | 0.3968 |
| R5-DP | 0.1341 | 0.3063 | 0.2097 | 0.0686 |
| R5-DV | 0.0585 | 0.4898 | 0.0896 | -0.1172 |
| | | | | |
| R6-DA | 0.0585 | 0.4303 | 0.0318 | 0.1557 |
| R6-DE | 0.3050 | 0.4303 | 0.0700 | 0.1318 |
| R6-DI | 0.0614 | 0.1908 | 0.0220 | 0.0834 |
| R6-DP | 0.3752 | 0.2467 | 0.0428 | 0.2336 |
| R6-DV | 0.0512 | 0.5498 | 0.0582 | 0.1442 |

Feature Set A (sub-set, CFS) , SVM

| Rater/predictor-dimension | Human rater MSE | Human rater correlation | Automatic predictor MSE | Automatic predictor correlation |
|---|---|---|---|---|
| R3-DA | 0.1843 | 0.6145 | 0.0646 | 0.4657 |
| R3-DE | 0.1813 | 0.3578 | 0.1800 | 0.2931 |
| R3-DI | 0.0766 | 0.4942 | 0.0818 | 0.3475 |
| R3-DP | 0.3233 | 0.4942 | 0.4277 | 0.2722 |
| R3-DV | 0.0725 | 0.4942 | 0.0953 | 0.3046 |
| | | | | |
| R5-DA | 0.1920 | 0.4942 | 0.1205 | 0.1584 |
| R5-DE | 0.3700 | 0.3228 | 0.1385 | 0.2077 |
| R5-DI | 0.1226 | 0.5104 | 0.0712 | 0.3921 |
| R5-DP | 0.1341 | 0.3063 | 0.2228 | 0.1310 |
| R5-DV | 0.0585 | 0.4898 | 0.0764 | -0.0174 |
| | | | | |
| R6-DA | 0.0585 | 0.4303 | 0.0297 | 0.2035 |
| R6-DE | 0.3050 | 0.4303 | 0.0796 | 0.2035 |
| R6-DI | 0.0614 | 0.1908 | 0.0240 | 0.0864 |
| R6-DP | 0.3752 | 0.2467 | 0.0479 | 0.2326 |
| R6-DV | 0.0512 | 0.5498 | 0.0587 | 0.2540 |

Feature Set A (sub-set CFS) + BOW features, SVM

| Rater/predictor-dimension | Human rater MSE | Human rater correlation | Automatic predictor MSE | Automatic predictor correlation |
|---|---|---|---|---|
| R3-DA | 0.1843 | 0.6145 | 0.0600 | 0.4140 |
| R3-DE | 0.1813 | 0.3578 | 0.2489 | 0.0397 |
| R3-DI | 0.0766 | 0.4942 | 0.0625 | 0.3471 |
| R3-DP | 0.3233 | 0.4942 | 0.3957 | 0.2985 |
| R3-DV | 0.0725 | 0.4942 | 0.1137 | 0.1689 |
| | | | | |
| R5-DA | 0.1920 | 0.4942 | 0.1127 | 0.1405 |
| R5-DE | 0.3700 | 0.3228 | 0.1976 | -0.0223 |
| R5-DI | 0.1226 | 0.5104 | 0.0650 | 0.3791 |
| R5-DP | 0.1341 | 0.3063 | 0.2225 | 0.0527 |
| R5-DV | 0.0585 | 0.4898 | 0.0766 | -0.0016 |
| | | | | |
| R6-DA | 0.0585 | 0.4303 | 0.0315 | 0.1722 |
| R6-DE | 0.3050 | 0.4303 | 0.0632 | 0.2753 |
| R6-DI | 0.0614 | 0.1908 | 0.0221 | 0.1366 |
| R6-DP | 0.3752 | 0.2467 | 0.0429 | 0.2339 |
| R6-DV | 0.0512 | 0.5498 | 0.0600 | 0.2079 |

Feature Set B (all) , SVM

| Rater/predictor-dimension | Human rater MSE | Human rater correlation | Automatic predictor MSE | Automatic predictor correlation |
|---|---|---|---|---|
| R3-DA | 0.1843 | 0.6145 | 0.0594 | 0.4063 |
| R3-DE | 0.1813 | 0.3578 | 0.2365 | 0.2698 |
| R3-DI | 0.0766 | 0.4942 | 0.0860 | 0.2407 |
| R3-DP | 0.3233 | 0.4942 | 0.3716 | 0.2968 |
| R3-DV | 0.0725 | 0.4942 | 0.1102 | 0.1321 |
| | | | | |
| R5-DA | 0.1920 | 0.4942 | 0.1138 | 0.1395 |
| R5-DE | 0.3700 | 0.3228 | 0.2009 | 0.1059 |
| R5-DI | 0.1226 | 0.5104 | 0.0716 | 0.3585 |
| R5-DP | 0.1341 | 0.3063 | 0.2336 | 0.3585 |
| R5-DV | 0.0585 | 0.4898 | 0.2336 | 0.0049 |
| | | | | |
| R6-DA | 0.0585 | 0.4303 | 0.0315 | 0.1870 |
| R6-DE | 0.3050 | 0.4303 | 0.0667 | 0.1915 |
| R6-DI | 0.0614 | 0.1908 | 0.0222 | 0.0787 |
| R6-DP | 0.3752 | 0.2467 | 0.0404 | 0.3610 |
| R6-DV | 0.0512 | 0.5498 | 0.0576 | 0.2246 |

Feature Set B (sub-set CFS) , SVM

| Rater/predictor-dimension | Human rater MSE | Human rater correlation | Automatic predictor MSE | Automatic predictor correlation |
|---|---|---|---|---|
| R3-DA | 0.1843 | 0.6145 | 0.0638 | 0.4584 |
| R3-DE | 0.1813 | 0.3578 | 0.1870 | 0.3263 |
| R3-DI | 0.0766 | 0.4942 | 0.0760 | 0.3464 |
| R3-DP | 0.3233 | 0.4942 | 0.3770 | 0.3391 |
| R3-DV | 0.0725 | 0.4942 | 0.0953 | 0.3168 |
| | | | | |
| R5-DA | 0.1920 | 0.4942 | 0.1138 | 0.1850 |
| R5-DE | 0.3700 | 0.3228 | 0.1543 | 0.1661 |
| R5-DI | 0.1226 | 0.5104 | 0.0749 | 0.4375 |
| R5-DP | 0.1341 | 0.3063 | 0.2468 | 0.0832 |
| R5-DV | 0.0585 | 0.4898 | 0.0627 | 0.0552 |
| | | | | |
| R6-DA | 0.0585 | 0.4303 | 0.0300 | 0.2139 |
| R6-DE | 0.3050 | 0.4303 | 0.0783 | 0.1461 |
| R6-DI | 0.0614 | 0.1908 | 0.0240 | 0.0885 |
| R6-DP | 0.3752 | 0.2467 | 0.0461 | 0.3283 |
| R6-DV | 0.0512 | 0.5498 | 0.0536 | 0.4082 |

Feature Set B (sub-set CFS) + BOW features, SVM

| Rater/predictor-dimension | Human rater MSE | Human rater correlation | Automatic predictor MSE | Automatic predictor correlation |
|---|---|---|---|---|
| R3-DA | 0.1843 | 0.6145 | 0.0590 | 0.4248 |
| R3-DE | 0.1813 | 0.3578 | 0.2439 | 0.0600 |
| R3-DI | 0.0766 | 0.4942 | 0.0687 | 0.3435 |
| R3-DP | 0.3233 | 0.4942 | 0.3963 | 0.2956 |
| R3-DV | 0.0725 | 0.4942 | 0.1153 | 0.1732 |
| | | | | |
| R5-DA | 0.1920 | 0.4942 | 0.1169 | 0.1117 |
| R5-DE | 0.3700 | 0.3228 | 0.1945 | -0.0388 |
| R5-DI | 0.1226 | 0.5104 | 0.0674 | 0.3416 |
| R5-DP | 0.1341 | 0.3063 | 0.2206 | 0.0410 |
| R5-DV | 0.0585 | 0.4898 | 0.0742 | 0.0318 |
| | | | | |
| R6-DA | 0.0585 | 0.4303 | 0.0314 | 0.1845 |
| R6-DE | 0.3050 | 0.4303 | 0.0635 | 0.2959 |
| R6-DI | 0.0614 | 0.1908 | 0.0221 | 0.1344 |
| R6-DP | 0.3752 | 0.2467 | 0.0421 | 0.2846 |
| R6-DV | 0.0512 | 0.5498 | 0.0588 | 0.2330 |

Feature Set C (all) , SVM

| Rater/predictor-dimension | Human rater MSE | Human rater correlation | Automatic predictor MSE | Automatic predictor correlation |
|---|---|---|---|---|
| R3-DA | 0.1843 | 0.6145 | 0.0594 | 0.4063 |
| R3-DE | 0.1813 | 0.3578 | 0.2382 | 0.2242 |
| R3-DI | 0.0766 | 0.4942 | 0.0860 | 0.2407 |
| R3-DP | 0.3233 | 0.4942 | 0.3694 | 0.2851 |
| R3-DV | 0.0725 | 0.4942 | 0.1104 | 0.1307 |
| | | | | |
| R5-DA | 0.1920 | 0.4942 | 0.1118 | 0.1586 |
| R5-DE | 0.3700 | 0.3228 | 0.1957 | 0.1147 |
| R5-DI | 0.1226 | 0.5104 | 0.0745 | 0.2984 |
| R5-DP | 0.1341 | 0.3063 | 0.2307 | 0.1184 |
| R5-DV | 0.0585 | 0.4898 | 0.0667 | 0.0049 |
| | | | | |
| R6-DA | 0.0585 | 0.4303 | 0.0315 | 0.1886 |
| R6-DE | 0.3050 | 0.4303 | 0.0659 | 0.2069 |
| R6-DI | 0.0614 | 0.1908 | 0.0218 | 0.1048 |
| R6-DP | 0.3752 | 0.2467 | 0.0404 | 0.3655 |
| R6-DV | 0.0512 | 0.5498 | 0.0578 | 0.2279 |

Feature Set C (sub-set, CFS) , SVM

| Rater/predictor-dimension | Human rater MSE | Human rater correlation | Automatic predictor MSE | Automatic predictor correlation |
|---|---|---|---|---|
| R3-DA | 0.1843 | 0.6145 | 0.0638 | 0.4584 |
| R3-DE | 0.1813 | 0.3578 | 0.1878 | 0.2701 |
| R3-DI | 0.0766 | 0.4942 | 0.0760 | 0.3464 |
| R3-DP | 0.3233 | 0.4942 | 0.3708 | 0.3448 |
| R3-DV | 0.0725 | 0.4942 | 0.1085 | 0.1705 |
| | | | | |
| R5-DA | 0.1920 | 0.4942 | 0.1120 | 0.2095 |
| R5-DE | 0.3700 | 0.3228 | 0.1523 | 0.1758 |
| R5-DI | 0.1226 | 0.5104 | 0.0802 | 0.3256 |
| R5-DP | 0.1341 | 0.3063 | 0.2498 | 0.1133 |
| R5-DV | 0.0585 | 0.4898 | 0.0627 | 0.0552 |
| | | | | |
| R6-DA | 0.0585 | 0.4303 | 0.0304 | 0.2057 |
| R6-DE | 0.3050 | 0.4303 | 0.0777 | 0.1497 |
| R6-DI | 0.0614 | 0.1908 | 0.0237 | 0.1046 |
| R6-DP | 0.3752 | 0.2467 | 0.0453 | 0.3708 |
| R6-DV | 0.0512 | 0.5498 | 0.0599 | 0.2741 |

Feature Set C (sub-set, CFS) + BOW features , SVM

Some preliminary experiments were performed using LSTM-RNN. The results are below those of SVR in most cases. However, the LSTM-RNN weren't tuned to the task yet, so this must be considered as preliminary experiments. Notably, however, is the increased performance for valence (esp. R3-DV).

The network is a unidirectional network with two hidden layers with LSTM memory blocks (1 cell each). The first layer has 70 blocks and the second layer 20 blocks. The network was trained using resilient propagation with backpropagation through time.

| Rater/predictor-dimension | Human rater MSE | Human rater correlation | Automatic predictor MSE | Automatic predictor correlation |
|---|---|---|---|---|
| R3-DA | 0.1843 | 0.6145 | 0.0734 | 0.3535 |
| R3-DE | 0.1813 | 0.3578 | 0.2193 | 0.2618 |
| R3-DI | 0.0766 | 0.4942 | 0.0619 | -0.0152 |
| R3-DP | 0.3233 | 0.4942 | 0.2296 | 0.1521 |
| R3-DV | 0.0725 | 0.4942 | 0.1044 | 0.3326 |
| | | | | |
| R5-DA | 0.1920 | 0.4942 | 0.1128 | 0.0677 |
| R5-DE | 0.3700 | 0.3228 | 0.1484 | 0.1870 |

| | | | | |
|---|---|---|---|---|
| R5-DI | 0.1226 | 0.5104 | 0.0671 | 0.3109 |
| R5-DP | 0.1341 | 0.3063 | 0.2017 | 0.0856 |
| R5-DV | 0.0585 | 0.4898 | 0.0592 | 0.0856 |
| | | | | |
| R6-DA | 0.0585 | 0.4303 | 0.0310 | 0.1190 |
| R6-DE | 0.3050 | 0.4303 | 0.1130 | 0.2821 |
| R6-DI | 0.0614 | 0.1908 | 0.0275 | 0.1868 |
| R6-DP | 0.3752 | 0.2467 | 0.0771 | 0.3076 |
| R6-DV | 0.0512 | 0.5498 | 0.0900 | 0.2353 |

Feature Set C (all), LSTM-RNN, 2 hidden layers (70/20).

**The interestDetector component**

Subject independent results have been computed on the TUM AVIC corpus in leave-one-speaker-group-out (LOSGO) setting. The same speaker groups as in (Schuller et al., 2009b) were used. Three levels of interest are distinguished (bored, neutral, interested). The classifier of choice is Support Vector Machines (SVM) with a polynomial kernel function of degree 1. We report weighted average class-wise recall (WA) and unweighted average class-wise recall (UA),

| Feature Set: | WA [%] | UA |
|---|---|---|
| Set A | 67.19 | 62.20 |
| Set B | 68.49 | 61.84 |
| Set C | 66.49 | 59.83 |

Leave-one-subject-group-out cross validation (5-fold), SVM, poly kernel, degree 1 (libSVM)

### 3.2.1 The EmotionFusion component

An evaluation of the fusion was performed on each rater, where the predicted output of the combined acoustic and linguistic/non-verbal emotion recognizer was added to the predicted output of the video based emotion recognizer. This final prediction sum was divided by two to obtain the average value. The individual results (both evaluated only during user speech regions – thus the big differences to the results in section 3.2.1) and the fusion can be seen in the following table:

| Rater/predictor-dimension | Video MSE | Video correlation | Audio MSE | Audio correlation | Fused MSE | Fused correlation |
|---|---|---|---|---|---|---|
| R3-DA | 0.0767 | 0.0159 | 0.0638 | 0.4584 | ***0.0692*** | ***0.3977*** |
| R3-DE | 0.2428 | -0.0045 | 0.1870 | 0.3263 | ***0.1930*** | ***0.2829*** |
| R3-DI | 0.1186 | 0.1785 | 0.0760 | 0.3464 | ***0.0902*** | ***0.3566*** |
| R3-DP | 0.4788 | -0.0528 | 0.3770 | 0.3391 | ***0.4300*** | ***0.2950*** |
| R3-DV | 0.1072 | -0.0223 | 0.0953 | 0.3168 | ***0.0962*** | ***0.2736*** |
| | | | | | | |
| R5-DA | 0.1127 | 0.1817 | 0.1138 | 0.1850 | ***0.1110*** | ***0.2174*** |
| R5-DE | 0.2288 | -0.0401 | 0.1543 | 0.1661 | ***0.1682*** | ***0.1617*** |
| R5-DI | 0.0727 | 0.1675 | 0.0749 | 0.4375 | ***0.0710*** | ***0.4650*** |
| R5-DP | 0.2246 | -0.0250 | 0.2468 | 0.0832 | ***0.2445*** | ***0.0460*** |
| R5-DV | 0.0535 | 0.0337 | 0.0627 | 0.0552 | ***0.0543*** | ***0.0510*** |
| | | | | | | |
| R6-DA | 0.0308 | 0.0516 | 0.0300 | 0.2139 | ***0.0285*** | ***0.2277*** |
| R6-DE | 0.0726 | -0.0567 | 0.0783 | 0.1461 | ***0.0796*** | ***0.0958*** |
| R6-DI | 0.0201 | -0.0567 | 0.0240 | 0.0885 | ***0.0220*** | ***0.2632*** |
| R6-DP | 0.0455 | 0.1789 | 0.0461 | 0.3283 | ***0.0473*** | ***0.3543*** |
| R6-DV | 0.0531 | 0.0498 | 0.0536 | 0.4082 | ***0.0546*** | ***0.4047*** |

Audio Feature Set B (sub-set, CFS) + BOW features, SVM ; late fusion with video prediction

# 4 License and availability

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 (including the feature extraction, keyword spotter, and bag-of-words component) is also available as a standalone open-source package (openSMILE) to the emotion research community (Eyben et al., 2010b).

## References

(Gunes & Pantic, 2010a) H. Gunes and M. Pantic: “Automatic, Dimensional and Continuous Emotion Recognition”, in International Journal of Synthetic Emotions, Vol. 1, No. 1, pp. 68-99, 2010.

(Gunes & Pantic, 2010b) H. Gunes and M. Pantic: “Dimensional Emotion Prediction from Spontaneous Head Gestures for Interaction with Sensitive Artificial Listeners”, in Proc. of Int. Conf. on Virtual Agents (IVA), pp. 371–377.

(Nicolaou et al., 2010a) M. A. Nicolaou, H. Gunes and M. Pantic: “Audio-visual Classification and Fusion of Spontaneous Affective Data in Likelihood Space”, in Proc. of ICPR 2010, the 20th IAPR Int. Conf. on Pattern Recognition, pp. 3695-3699.

(Nicolaou et al., 2010b) M. A. Nicolaou, H. Gunes and M. Pantic: “Automatic Segmentation of Spontaneous Data using Dimensional Labels from Multiple Coders” in Proc. of LREC Int. Workshop on Multimodal Corpora, pp. 43-48.

(Schuller et al., 2009a) B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, H. Konosu: *"Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application"*, Image and Vision Computing Journal (IVCJ), invited for the Special Issue on "Visual and Multimodal Analysis of Human Spontaneous Behavior", ELSEVIER, Vol. 27, Issue 12, pp. 1760-1774, November 2009.

(Schuller et al., 2009b) B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth: *"Acoustic Emotion Recognition: A Benchmark Comparison of Performances"*, Proc. Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, pp. 552-557, Merano, Italy, 13.-17.12.2009.

(Schuller et al., 2010a) B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, S. Narayanan: *"The INTERSPEECH 2010 Paralinguistic Challenge"*, to appear in Proc. INTERSPEECH 2010, ISCA, Makuhari, Japan, 26.-30.09.2010

(Eyben et al., 2010a) 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"*, Journal on Multimodal User Interfaces (JMUI), Special Issue on "Real-Time Affect Analysis and Interpretation: Closing the Affective Loop in Virtual Agents and Robots", Springer, Berlin/Heidelberg, ISSN 1783-7677, vol. 3, no. 1-2, pp. 7-12, March 2010

(Eyben et al., 2010b) F. Eyben, M. Wöllmer, B. Schuller: *"openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor"*, to appear in Proc. ACM Multimedia, Firenze, Italy, 25.-29.10.2010.