AVEC 2011 – The First International Audio/Visual Emotion Challenge*

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

[1]Technische Universität München
Institute for Human-Machine Communication, Munich, Germany
[2]Imperial College London, Intelligent Behaviour Understanding Group, London, UK
[3]Queen's University, School of Psychology, Belfast, BT7 1NN, UK
[4]Twente University, EEMCS, Twente, The Netherlands
http://sspnet.eu/avec2011

Abstract. The Audio/Visual Emotion Challenge and Workshop (AVEC 2011) is the first competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual and audiovisual emotion analysis, with all participants competing under strictly the same conditions. This paper first describes the challenge participation conditions. Next follows the data used – the SEMAINE corpus – and its partitioning into train, development, and test partitions for the challenge with labelling in four dimensions, namely activity, expectation, power, and valence. Further, audio and video baseline features are introduced as well as baseline results that use these features for the three sub-challenges of audio, video, and audiovisual emotion recognition.

Keywords: Audiovisual Emotion Recognition, Speech Emotion Recognition, Facial Expression Analysis, Challenge

1 Introduction

The Audio/Visual Emotion Challenge and Workshop (AVEC 2011) is the first competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual, and audiovisual emotion analysis, with all participants competing under strictly the same conditions. The goal of the challenge is to provide a common benchmark test partition 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 behaviour in large volumes of un-segmented,

* The authors would like to thank the sponsors of the challenge, the Social Signal Processing Network (SSPNet) and the HUMAINE Association. The responsibility lies with the authors.

non-prototypical and non-preselected data as this is exactly the type of data that both multimedia retrieval and human-machine/human-robot communication interfaces have to face in the real world. As the benchmark database the SEMAINE database of naturalistic dialogues will be used. Three Sub-Challenges are addressed:

- In the *Audio Sub-Challenge*, exclusively audio feature information is used at the word level.
- In the *Video Sub-Challenge*, exclusively video feature information is used at the frame level.
- In the *Audiovisual Sub-Challenge*, audiovisual feature information is used at the word level.

Four classification problems need to be solved for Challenge participation: the originally continuous dimensions Activity (arousal), Expectation, Power, and Valence were redefined as binary classification tasks by testing at every frame whether they were above or below mean. The Challenge competition measure is classification accuracy averaged over all four dimensions. All Sub-Challenges allow contributors to find their own features and use them with their own classification algorithm. However, standard feature sets (for audio and video separately) are given that may be used. The labels of the test partition remain unknown to the participants, and participants have to stick to the definition of training, development, and test partition. They may freely report on results obtained on the development partition, but have only a limited number of five trials per Sub-Challenge to submit their results on the test partition, whose labels are unknown to them. To ensure that unimodal results on test are really based on this modality, the test partition has been further split into three test sub-partitions, one for each Sub-Challenge, with either exclusively the audio or video or, for the audiovisual task, both tracks available.

To be eligible to participate in the challenge, every entry has to be accompanied by a paper presenting the results and the methods that created them, which will undergo peer-review. Only contributions with an accepted paper will be eligible for the Challenge participation. The organisers preserve the right to re-evaluate the findings, but will not participate themselves in the Challenge. Participants are encouraged to compete in all Sub-Challenges.

We next introduce the Challenge corpus (Sec. 2) and labels (Sec. 3), then audio and visual baseline features (Sec. 4), and baseline results (Sec. 5), before concluding in Sec. 6.

2 SEMAINE Database

The SEMAINE corpus [11], freely available for scientific research purposes from http://semaine-db.eu, was recorded to study natural social signals that occur in conversations between humans and artificially intelligent agents, and to collect data for the training of the next generation of such agents. The scenario used is called the Sensitive Artificial Listener (SAL) [4]. It involves a user interacting

with emotionally stereotyped "characters" whose responses are stock phrases keyed to the user's emotional state rather than the content of what (s)he says.

For the recordings, the participants are asked to talk in turn to four emotionally stereotyped characters. These characters are Prudence, who is even-tempered and sensible; Poppy, who is happy and outgoing; Spike, who is angry and confrontational; and Obadiah, who is sad and depressive.

Video was recorded at 49.979 frames per second at a spatial resolution of 780 x 580 pixels and 8 bits per sample, while audio was recorded at 48 kHz with 24 bits per sample. To accommodate research in audio-visual fusion, the audio and video signals were synchronised with an accuracy of 25 μs using the system developed by Lichtenauer et al. [10].

The part of the database used in this challenge consists of 24 recordings, with approximately 4 character conversation sessions per recording. This part was split into three partitions for the AVEC challenge: a training, development, and test partition each consisting of 8 recordings. Because the number of character conversations varies somewhat between recordings, the number of sessions (and thus audio and video files) is different per set: The training partition contains 31 sessions, while the development and test partitions contain 32 sessions. Table 1 shows the distribution of data in sessions, video frames, and words for each partition. A separate website was set up for the AVEC 2011 competition data[1].

Table 1. Overview of dataset make-up per partition

# / (h:m:s) / [ms]	**Train**	**Development**	**Test**	**Total**
Sessions	31	32	32	95
Frames	501 277	449 074	407 772	1 358 123
Words	20 183	16 311	13 856	50 350
Total duration	2:47:10	2:29:45	2:15:59	7:32:54
Avg. word duration	262	276	249	263

3 Challenge Labels

For the challenge, we selected the affective dimensions for which all character interactions of the Solid-SAL part are annotated by at least two raters. These are the dimensions Activity, Expectation, Power, and Valence, which are all well established in the psychological literature. An influential recent study [7] argues that these four dimensions account for most of the distinctions between everyday emotion categories.

Activity is the individual's global feeling of dynamism or lethargy. It subsumes mental activity as well as physical preparedness to act as well as overt activity. Expectation (Anticipation) also subsumes various concepts that can

[1] `http://avec2011-db.sspnet.eu/`

be separated as expecting, anticipating, being taken unaware. Again, they point to a dimension that people find intuitively meaningful, related to control in the domain of information. The POWER (Dominance) dimension subsumes two related concepts, power and control. However, people's sense of their own power is the central issue that emotion is about, and that is relative to what they are facing. VALENCE is an individual's overall sense of "weal or woe": Does it appear that on balance, the person rated feels positive or negative about the things, people, or situations at the focus of his/her emotional state?

Table 2. Overview of class balance: fraction of positive instances over total instances of video frames and words in training and test partition.

Ratio	ACTIVITY	EXPECTATION	POWER	VALENCE
Frames training	0.466	0.455	0.512	0.547
Frames development	0.555	0.397	0.588	0.636
Words training	0.496	0.409	0.560	0.554
Words development	0.581	0.334	0.670	0.654

All interactions were annotated by 2 to 8 raters, with the majority annotated by 6 raters: 68.4 % of interactions were rated by 6 raters or more, and 82 % by 3 or more. The raters annotated the four dimensions in continuous time and continuous value using a tool called FeelTrace [3], and the annotations are often called *traces*. This resulted in a set of trace vectors $\{\mathbf{v}_i^a, \mathbf{v}_i^e, \mathbf{v}_i^p, \mathbf{v}_i^v\} \in \mathbb{R}$ for every rater i and dimension a (ACTIVITY), e (EXPECTATION), p (POWER), and v (VALENCE). To attain binary labels, we first computed the average value of each dimension over all raters, resulting in a set of continuous time, real valued variables $\{\hat{\mathbf{v}}^a, \hat{\mathbf{v}}^e, \hat{\mathbf{v}}^p, \hat{\mathbf{v}}^v\} \in \mathbb{R}$. We then computed the mean of these average ratings over all interactions in the dataset, resulting in the scalar values $\{\mu^a, \mu^e, \mu^p, \mu^v\} \in \mathbb{R}$. The binary labels $\{\mathbf{y}^a, \mathbf{y}^e, \mathbf{y}^p, \mathbf{y}^v\} \in \{\pm 1\}$ are then found by thresholding $\hat{v}_t^j > \mu^j$ for each dimension j at every frame t.

For the *Video Sub-Challenge*, the original traces are binned in temporal units of the same duration as a single frame (i. e., 1/49.979 seconds), resulting in a binary label per frame. For the audio, the traces are binned over the duration of the words uttered by the user, resulting in a single binary label per word. The word timings were obtained by running an HMM-based speech recogniser in forced alignment mode on the manual transcripts of the interactions. The recogniser uses tied-state cross-word triphone left-right (linear) HMM models with 3 emitting states and 16 Gaussian mixture components per state. Monophones with 1 Gaussian mixture component per state were bootstrapped on all available speech data (user and operator) of the SEMAINE corpus. The tied-state triphone models were created from these initial monophone models by decision tree based state clustering and the number of Gaussian mixture components was increased to 16 in four iterations of successive mixture doubling. In order to use accessible standard tool kits for maximum reproducibility of results, the

Hidden Markov Toolkit (HTK) [18] was used to train the models and create the alignments.

Tables 1 and 2 provide an overview of the AVEC 2011 competition dataset. Table 1 lists the number of interactions per data partition, and the number of video instances (i. e., frames) and audio/audio-visual instances (i. e., words). It also reports the average word duration, in milliseconds. Table 2 lists the fraction of positive instances per partition and per dimension. It shows that the data is fairly balanced – owed to the design choice of the two classes positive/negative being defined as above/below mean. This led to the use of the classification accuracy (weighted average accuracy, WA) as the performance measure in this Challenge.

Some of the dimensions are highly correlated. For example, in the training and development partitions, at the frame-level, expectation and power are negatively correlated by a factor of 0.373. The full correlation matrices for both word-level and frame-level labels are given in Table 3. All correlations have a p-value $<< 0.01$.

Table 3. Correlation coefficients (CC) for the dimensions at the word and frame level. (E) denotes Expectation, (P) Power, and (V) Valence.

CC	**Word level**			**Frame level**		
[%]	E	P	V	E	P	V
Activation	-3.2	22.4	20.7	-3.2	24.5	24.9
Expectation		-35.8	-10.4		-37.3	-7.7
Power			29.7			29.6

4 Baseline Features

In the following sections we describe how the publicly available baseline feature sets are computed for either the audio or the video data. Participants could use these feature sets exclusively or in addition to their own features.

4.1 Audio Features

In this Challenge, an extended set of features with respect to the INTERSPEECH 2009 Emotion Challenge (384 features) [13] and INTERSPEECH 2010 Paralinguistic Challenge (1 582 features) [14] is given to the participants, again using the freely available open-source Emotion and Affect Recognition (openEAR) [5] toolkit's feature extraction backend openSMILE [6].

The audio baseline feature set consists of 1 941 features, composed of 25 energy and spectral related low-level descriptors (LLD) x 42 functionals, 6 voicing related LLD x 32 functionals, 25 delta coefficients of the energy/spectral LLD x

Table 4. 31 low-level descriptors.

Energy & spectral (25)
loudness (auditory model based), zero crossing rate, energy in bands from 250 – 650 Hz, 1 kHz – 4 kHz, 25 %, 50 %, 75 %, and 90 % spectral roll-off points, spectral flux, entropy, variance, skewness, kurtosis, psychoacousitc sharpness, harmonicity, MFCC 1-10
Voicing related (6)
F_0 (sub-harmonic summation (SHS) followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), jitter (delta: "jitter of jitter"), logarithmic Harmonics-to-Noise Ratio (logHNR)

23 functionals, 6 delta coefficients of the voicing related LLD x 19 functionals, and 10 voiced/unvoiced durational features. Details for the LLD and functionals are given in tables 4 and 5 respectively. The set of LLD covers a standard range of commonly used features in audio signal analysis and emotion recognition. The functional set has been based on similar sets, such as the one used for the INTERSPEECH 2011 Speaker State Challenge [15], but has been carefully reduced to avoid LLD/functional combinations that produce values which are constant, contain very little information, and/or high amount of noise.

The audio features are computed on short episodes of audio data of variable duration. To wit, one instance is recorded for every word uttered by the user in a SAL interaction. Since the timings of the word boundaries were estimated by a speech recogniser with forced alignment using the manually created transcripts of the interactions, it is possible that some of the word boundaries are calculated incorrectly. In particular, some of the words were found to be so short that it is impossible to compute the audio features. To alleviate this problem, for words that were found to be too short we artificially changed the start and end time of the word to attain a segment with a minimum length of 0.25 s. The actual annotated word thereby was placed in the centre of this segment.

4.2 Video Features

The bulk of the features extracted from the video streams of the character interactions are computed by dense local appearance descriptors. These descriptors are most informative if they are applied to frontal faces of uniform size. Since the head pose and distance to the camera varies over time in the SEMAINE recordings, we detect the locations of the eyes to help remove this variance. The information describing the position and pose of the face and eyes are valuable for detecting the dimensional affect in themselves and are thus added to the set of video features, too.

Table 5. Set of all 42 functionals. [1]Not applied to delta coefficient contours. [2]For delta coefficients the mean of only positive values is applied, otherwise the arithmetic mean is applied. [3]Not applied to voicing related LLD.

Statistical functionals (23)
(positive[2]) arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, 1 %, 99 % percentile, percentile range 1 %–99 %, percentage of frames contour is above: minimum + 25%, 50%, and 90 % of the range, percentage of frames contour is rising, maximum, mean, minimum segment length[3], standard deviation of segment length[3]
Regression functionals[1] (4)
linear regression slope, and corresponding approximation error (linear), quadratic regression coefficient a, and approximation error (linear)
Local minima/maxima related functionals[1] (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances, amplitude mean of maxima, amplitude mean of minima, amplitude range of maxima
Other[1,3] (6)
LP gain, LPC 1 – 5

To obtain the face position, we employ another open-source available implementation – OpenCV's Viola & Jones face detector. This returns a four-valued face position and size descriptor, to wit, the x and y position of the top-left corner of the face area, and the width and height of the face area. The height and width output of this detector is rather unstable: Even in a video in which a face hardly moves the values for the height and width vary significantly (approximately 5 % standard deviation). Also, the face detector does not provide any information about the head pose. To refine the detected face region, and allow the appearance descriptor to correlate better with the shown expression instead of with variability in head pose and face detector output, we proceed with detection of the locations of the eyes. This is again done with the OpenCV implementation of a Haar-cascade object detector, trained for either a left or a right eye. After the left eye location p_l and right eye location p_r are determined, the image is rotated to set the angle α, defined as the angle between the line connecting the eyes and the horizontal, to be 0 degrees, scaled to make the distance between the eye locations 100 pixels, and then cropped to be 200 by 200 pixels, with p_r at position $\{p_r^x, p_r^y\} = \{80, 60\}$ to obtain the registered face image. The eye locations are included as part of the video features provided for candidates.

As dense local appearance descriptors we chose to use uniform Local Binary Patterns (LBP) [12]. They have been used extensively for face analysis in recent years, e. g., for face recognition [1], emotion detection [16], or detection of facial muscle actions (FACS Action Units) [9]. They were also used as the baseline features for the recently held challenge on facial expression recognition and analysis (FERA 2011, [17]). Consisting of 8 binary comparisons per pixel, they are fast

to compute. By employing uniform LBPs instead of full LBPs and aggregating the LBP operator responses in histograms taken over regions of the face, the dimensionality of the features is rather low (59 dimensions per image block). In our baseline method and feature extraction implementation we divided the registered face region into 10 x 10 blocks, resulting in 5 900 features.

Not provided, but used in the baseline method, are the head tilt α, and the distance between the eyes in the original image $d = ||p_r - p_l||^2$. p_r and p_l thereby are the position of the right and left eyes, accordingly.

5 Challenge Baselines

For transparency and easy reproducibility, we use Support Vector Machines (SVM) classification without feature selection. For the *Audio Sub-Challenge*, we used SVMs with linear Kernel, Sequential Minimal Optimization (SMO) for learning, and optimised the complexity on the development partition of the corpus. The SMO implementation in the WEKA toolkit is used [8]. For the *Video Sub-Challenge*, a SVM with a radial basis function (RBF) kernel was used instead implemented in the LibSVM tool [2]. For the *Audiovisual Sub-Challenge*, we first obtained predictions of the audio and video classifiers separately on both the development set and the audio-visual test set, in terms of posterior probabilities per word. We then fused the two modalities by concatenating the two posterior probabilities and trained a linear SVM on the development set data.

Because of the large number of data (over 1.3 million frames) and relatively high feature dimensionality (5 908 features per frame), due to memory constraints it is impossible to train a model using all data on a desktop PC. Instead, we sampled 1 000 frames from the training partition and 1 000 frames from the development partition. These were evenly divided over training/development videos (e. g., $k = \lfloor 1\,000/31 \rfloor = 32$ for training and $k = \lfloor 1\,000/32 \rfloor = 31$ for test). Within a video of n frames length, instances sampled had index $\lfloor i * n/k \rfloor$ with $i \in \{1 \ldots k\}$. For audio, training with the full data set is possible, however, to reproduce similar conditions as for video, we sub-sampled the data by using only every third word from the training and development partition.

Results per Sub-Challenge are given in Table 6 for training on the train partition and testing on the development partition – this can be freely done by participants – as well as for training on the unification of the training and the development partition and testing on the test partition sub-set for each Sub-Challenge. These results can be uploaded five times by the participants. To allow a comparison between the different approaches, we have provided the results of using only audio, only video, and using both for the *Audiovisual Sub-Challenge*.

The baseline results show that, while reasonable results are attained on the development set, a fair amount of overfitting appears to occur. Note that some scores are below chance. Further note that, the test sets for the *Audio* and the *Video Sub-Challenge* are different, and their results can thus not be used to compare the performance of audio vs. video based methods.

For the *Audiovisual Sub-Challenge*, we show the results of using only the audio baseline classifiers, video baseline classifiers, and fusing the audio and video modalities (last three rows of Table 6). The results show that, on the audio-visual test set, video has a better performance for all dimensions except for Expectation. Fusing the audio and video results shows mixed results: for Power and Valence we observe a marked improvement. For Activity and expectation the results are lower than the maximum of either audio or video, though.

Table 6. Baseline results per Sub-Challenge: (A) denotes *Audio*, (V) *Video*, and (AV) *Audiovisual Sub-Challenge*. WA stands for weighted accuracy, UA for unweighted accuracy. The mean over the four dimensions in the last coloumn is the overall competition measure used to rank participants in the three Sub-Challenges as typeset in boldface.

Accuracy	Activity		Expectation		Power		Valence		**Mean**
[%]	**WA**	UA	**WA**	UA	**WA**	UA	**WA**	UA	**WA**
Audio Sub-Challenge									
Development	63.7	64.0	63.2	52.7	65.6	55.8	58.1	52.9	62.7
Test	**55.0**	57.0	**52.9**	54.5	**28.0**	49.1	**44.3**	47.2	**45.1**
Video Sub-Challenge									
Development	60.2	57.9	58.3	56.7	56.0	52.8	63.6	60.9	59.5
Test	**42.2**	52.5	**53.6**	49.3	**36.4**	37.0	**52.5**	51.2	**46.2**
Audiovisual Sub-Challenge									
Test (A)	51.2	51.2	59.2	49.5	52.7	45.9	55.8	46.5	54.7
Test (V)	77.1	77.2	36.8	45.5	53.7	52.9	60.8	47.6	57.1
Test (AV)	**67.2**	67.2	**36.3**	48.5	**62.2**	50.0	**66.0**	49.2	**57.9**

6 Conclusion

We introduced AVEC 2011 – the first combined open Audio/Visual Emotion Challenge, its conditions, data, baseline features and results. By intention, we had preferred open-source software and highest transparency and realism for the baselines by refraining from feature space optimisation and optimising on test data. These baseline results indicate that this is a challenging problem indeed: On the test partitions, the official baseline sur-passes chance-level on average over binarised dimensions only for the *Audiovisual Sub-Challenge*.

Following the Challenge, we plan to combine all participants' results of the challenge by voting or meta-learning.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12), 2037–2041 (2006)

2. Chang, C.C., Lin, C.J.: LibSVM: a library for support vector machines (2001), software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm
3. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: Feeltrace: An instrument for recording perceived emotion in real time. In: Proc. of ISCA Workshop on Speech and Emotion. pp. 19–24. Belfast, UK (2000)
4. Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., Heylen, D.: The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In: LREC Workshop on Corpora for Research on Emotion and Affect. pp. 1–4. ELRA, Paris, France (2008)
5. Eyben, F., Wöllmer, M., Schuller, B.: openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In: Proc. ACII. pp. 576–581. Amsterdam, The Netherlands (2009)
6. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proc. ACM Multimedia (MM). pp. 1459–1462. Florence, Italy (2010)
7. Fontaine, J., K.R., S., Roesch, E., Ellsworth, P.: The world of emotions is not two-dimensional. Psychological science 18(2), 1050 – 1057 (2007)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations 11(1) (2009)
9. Jiang, B., Valstar, M., Pantic, M.: Action unit detection using sparse appearance descriptors in space-time video volumes. In: Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition. pp. 314–321. Santa Barbara, USA (2011)
10. Lichtenauer, J., Valstar, M.F., Shen, J., Pantic, M.: Cost-effective solution to synchornized audio-visual capture using multiple sensors. Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance pp. 324–329 (2009)
11. McKeown, G., Valstar, M., Pantic, M., Cowie, R.: The SEMAINE corpus of emotionally coloured character interactions. Proc. IEEE Int. Conf. Multimedia & Expo pp. 1–6 (2010)
12. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
13. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 Emotion Challenge. In: Proc. INTERSPEECH 2009. pp. 312–315. Brighton, UK (2009)
14. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: The INTERSPEECH 2010 Paralinguistic Challenge. In: Proc. INTERSPEECH 2010. pp. 2794–2797. Makuhari, Japan (2010)
15. Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J.: The INTERSPEECH 2011 Speaker State Challenge. In: Proc. INTERSPEECH 2011. ISCA, Florence, Italy (2011)
16. Shan, C., Gong, S., Mcowan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing 27(6), 803–816 (2009)
17. Valstar, M., Jiang, B., Mehu, M., Pantic, M., Scherer, K.: The first facial expression recognition and analysis challenge. Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition pp. 921–926 (2011)
18. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK book (v3.4). Cambridge University Press, Cambridge, UK (2006)