AVEC 2012: the continuous audio/visual emotion challenge

Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, Maja Pantic

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. 2012. "AVEC 2012: the continuous audio/visual emotion challenge." In *Proceedings of the 14th ACM International Conference on Multimodal Interaction - ICMI '12, October 2012, Santa Monica, CA, USA*, edited by Louis-Philippe Morency, Dan Bohus, Hamid Aghajan, Justine Cassell, Anton Nijholt, and Julien Epps, 449–56. New York, NY: ACM Press. https://doi.org/10.1145/2388676.2388776.

AVEC 2012 – The Continuous Audio/Visual Emotion Challenge

Björn Schuller
JOANNEUM RESEARCH Forschungsgesellschaft mbH*
DIGITAL - Institute for Information and Communication Technologies
Graz, Austria

Michel Valstar
University of Nottingham
Mixed Reality Lab
Nottingham, UK

Florian Eyben
Technische Universität München
Institute for Human-Machine Communication
Munich, Germany

Roddy Cowie
Queen's University
School of Psychology
Belfast, UK

Maja Pantic†
Imperial College London
Intelligent Behaviour Understanding Group
London, UK

ABSTRACT

We present the second Audio-Visual Emotion recognition Challenge and workshop (AVEC 2012), which aims to bring together researchers from the audio and video analysis communities around the topic of emotion recognition. The goal of the challenge is to recognise four continuously valued affective dimensions: arousal, expectancy, power, and valence. There are two sub-challenges: in the Fully Continuous Sub-Challenge participants have to predict the values of the four dimensions at every moment during the recordings, while for the Word-Level Sub-Challenge a single prediction has to be given per word uttered by the user. This paper presents the challenge guidelines, the common data used, and the performance of the baseline system on the two tasks.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

*The author is further affiliated with Technische Universität München, Munich, Germany

†The author is further affiliated with Twente University, EEMCS, Twente, The Netherlands.

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.

Keywords

Affective Computing, Emotion Recognition, Speech, Facial Expression, Challenge

1. INTRODUCTION

Dimensional affect recognition aims to improve the understanding of human affect by modelling affect as a small number of continuously valued, continuous time signals. Compared to the more limited categorical emotion description (e.g. six basic emotions), and for contemporary computational modelling techniques intractable appraisal theory, dimensional affect modelling has the benefit of being able to: a. encode small changes in affect over time, and b. distinguish between many more subtly different displays of affect, while remaining within the reach of current signal processing and machine learning capabilities.

The 2012 Audio-Visual Emotion Challenge and Workshop (AVEC 2012) will be the second competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, video 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 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 behaviour in large volumes of un-segmented, 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.

Following up from AVEC 2011 [16], which used a categorical description of affect in terms of low or high arousal,

expectancy, power, and valence, AVEC 2012 aims to accelerate research in automatic continuous affect recognition from audio and/or video. Whereas in AVEC 2011 the dimensional affect recognition problem was essentially reduced to a binary classification problem, for AVEC 2012 it is more naturally posed as a regression problem, and can thus be considered to be both more challenging and rewarding. Another difference between the first and second AVEC is that the first had separate categories for the audio, video, and audio-visual communities to compete in. Instead, for the present challenge we encourage participants to leverage both modalities.

We are calling for teams to participate in emotion recognition from acoustic audio analysis, linguistic audio analysis, video analysis, or any combination of these. As benchmarking database the SEMAINE database of naturalistic video and audio of human-agent interactions [11] will be used, which contains labels for the four target affect dimensions, amongst others. Emotion will have to be recognised in terms of continuous time, continuous valued dimensional affect in the dimensions arousal, expectation, power and valence. The database was recorded as part of the SEMAINE Sensitive Artificial Listener project [13]. Previously a subset of this data was used to perform continuous affect recognition on the dimensions Arousal, Expectancy, Intensity, Power, and Valence [7].

Two Sub-Challenges are addressed in AVEC 2012:

- The *Fully Continuous Sub-Challenge (FCSC)* involves fully continuous affect recognition, where the level of affect has to be predicted for every moment of the recording.

- The *Word-Level Sub-Challenge (WLSC)* requires participants to predict the level of affect at word-level, where a single value of affect has to be predicted per word, and only when the user is speaking.

Four regression problems need to be solved for Challenge participation: the continuous dimensions Arousal, Expectation, Power, and Valence. The Challenge competition measure is cross correlation averaged over all character interactions and all four dimensions.

Both Sub-Challenges allow contributors to find their own features to use with their regression algorithm. However, standard feature sets are provided (for audio and video separately), which participants are free to use. 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 are limited to five trials per Sub-Challenge in submitting their results on the test partition.

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 Challenge participation. The organisers preserve the right to re-evaluate the findings, but will not participate in the Challenge themselves.

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.

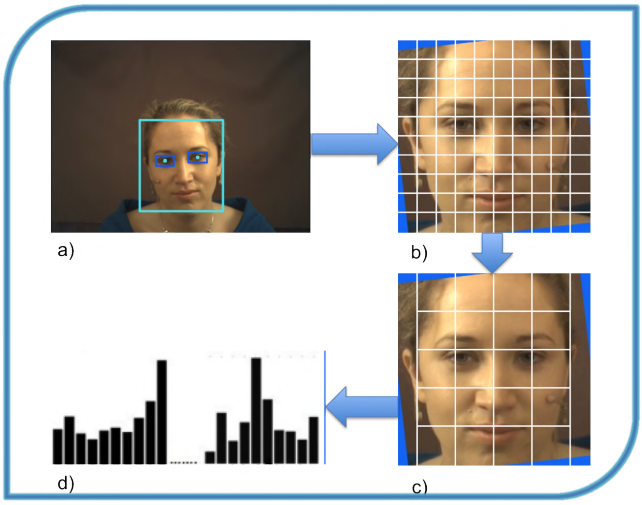

Figure 1: Video feature extraction overview: a) detection of face and eyes b) face normalised based on eye locations, divided in 10 x 10 blocks from which LBP features are extracted c). grid reduced to 5 x 4 to achieve dimensionality reduction d) histograms of separate blocks concatenated into single histogram.

2. SEMAINE DATABASE

The challenge uses the SEMAINE corpus [11] as the source of data. This database 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. It is freely available for scientific research purposes from http://semaine-db.eu. The scenario used in the recordings is called the Sensitive Artificial Listener (SAL) technique [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].

In this challenge the 24 recordings of the Solid-SAL part of the database were used, in which the characters are role-played by human operators. There are usually 4 character conversation sessions per recording. This Solid-SAL part was split into three partitions for the AVEC challenge: a training, development, and test partition each consisting of 8 recordings of 8 different users. 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 ses-

Table 1: Mapping between AVEC 2012 data and corresponding SEMAINE sessions

AVEC train	SEMAINE	AVEC devel	SEMAINE
1	2	1	8
2	3	2	9
3	4	3	10
4	5	4	11
5	29	5	19
6	30	6	20
7	31	7	21
8	40	8	22
9	41	9	34
10	42	10	35
11	43	11	36
12	58	12	37
13	59	13	46
14	60	14	47
15	61	15	48
16	70	16	49
17	71	17	82
18	72	18	83
19	73	19	84
20	76	20	85
21	77	21	94
22	78	22	95
23	79	23	96
24	88	24	97
25	89	25	112
26	90	26	113
27	91	27	114
28	106	28	115
29	107	29	131
30	108	30	132
31	109	31	133
		32	134

sions. Table 2 shows the distribution of data in sessions, video frames, and words for each partition.

The data for the AVEC 2012 was re-organised according to the partitions specified above. To allow researchers to relate the challenge data to the original data, we have provided a mapping in Table 1 for the training and development partitions. The data of the test partition is not available from the SEMAINE web portal. A separate website (`http://AVEC2011-db.sspnet.eu/`) was set up to distribute the AVEC 2012 competition data, which includes pre-computed audio and video features.

3. CHALLENGE LABELS

The affective dimensions used in the challenge were selected based on the available ratings. Dimensions for which all character interactions of the Solid-SAL part are annotated by at least two raters were included. These are the dimensions Arousal, Expectation, Power, and Valence, which are all well established in the psychological literature. An influential recent study [8] argues that these four dimensions account for most of the distinctions between everyday emotion categories.

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

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*. The annotation process 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 (Arousal), e (Expectation), p (Power), and v (Valence). The original traces are binned in temporal units of the same duration as a single video frame (i. e., 1/49.979 seconds). The labels for Arousal, Power, and Valence lie in the range $[-1, 1]$, and the labels for Expectation in the range $[0, 100]$.

Due to different ways in which video readers such as quicktime deal with keyframes in the H264 codecs, there are very small differences between the label vector lengths and the actual duration of the recordings. The difference is however very small, with the difference between the number of label instances and actual number of frames in the range of 0-0.04%. Participants are expected to deal with this by clipping/padding the label data as necessary.

In contrast with the first Audio-Visual Emotion recognition challenge (AVEC 2011), we use the fully continuous values as the challenge labels. So, while AVEC 2011 was a binary classification task, AVEC 2012 is a regression task. To obtain a single label per dimension rather than a set of labels with cardinality equal to the number of raters, we take the simple mean over the raters.

Two modes of label segmentation are given, one for each sub-challenge. For the FCSC, the labels that are binned per video frame are used. For the WLSC the traces are binned over the duration of the words uttered by the user, resulting in a single continuous value label per word. To segment the labels for the WCSC the word alignments available with the SEMAINE database were used. These 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) [19] was used to train the models

Table 2: 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

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

and create the alignments. These word timings are provided with the challenge data.

Table 2 lists the number of interactions per data partition, and the number of FCSC instances (i. e., frames) and WLSC instances (i. e., words). It also reports the average word duration, in milliseconds.

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 $<< 0.01$.

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, as was the case for AVEC 2011, an extended set of features with respect to the INTERSPEECH 2009 Emotion Challenge (384 features) [14] and INTERSPEECH 2010 Paralinguistic Challenge (1 582 features) [15] 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]. In contrast to AVEC 2011, the feature set was reduced by 100 features that were found to carry very little information, as they were zero or close to zero most of the time.

Thus, the AVEC 2012 audio baseline feature set consists of 1 841 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 19 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.

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, followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), jitter (delta: "jitter of jitter"), logarithmic Harmonics-to-Noise Ratio (logHNR)

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[1,3], standard deviation of segment length[1,3]
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

The audio features are computed on short episodes of audio data. Depending on the sub-challenge, these episodes are either whole words (for the WLSC) or 2 second sliding windows (for the FCSC). For the WLSC, one audio feature vector per word is extracted, while for the FCSC feature vectors are extracted at 0.5 second intervals, but only during speech (this includes non-linguistic vocalisations such as sighs and laughs). The first FCSC audio feature vector for every word is timed at the beginning of the word, and this time is thus not necessarily a multiple of 0.5 seconds. They thus do not align perfectly with the FCSC video feature vectors (see below).

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 dense local appearance descriptions. The descriptors that generate these features are most effective if they are applied to frontal faces of uniform size. Since the head pose and distance to the camera vary over time in the SEMAINE recordings, we detect the locations of the eyes to help reduce this variance. The information describing the position and pose of the face and eyes are in themselves valuable for recognising the dimensional affect and are thus included with the set of video features together with the appearance descriptors.

To obtain the face position, we employ another open-source implementation – OpenCV's Viola & Jones face detector. This returns a four-valued face position and size descriptor, to wit, the position of the top-left corner of the detected face area (f_x, f_y), followed by its width f_w and height f_h. 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). The face detector also doesn't 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. Let us define the detected left and right eye locations as p_l respectively p_r, and the the angle between the line connecting p_l and p_r, and the horizontal as α. The registered image is then obtained by rotating it to set $\alpha = 0$ degrees, scaled to make the distance between the eye locations $||p_l - p_r|| = 100$ 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 [17], 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, [18]). 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. The LBP histograms of the blocks are concatenated in lexicographic order resulting in the set F of 5 900 features. The video features are thus stored as follows: $\{f_x, f_y, f_w, f_h, p_r, p_l, F\}$, 5 908 features in total.

5. CHALLENGE BASELINES

For transparency and reproducibility, we use Support Vector Machine regression (SVR) without feature selection. We used SVRs with Histogram Intersection Kernels, Sequential Minimal Optimization (SMO) for learning. For evaluation on the test set (which is what the challenge scores are based on), all relevant parameters were optimised on the development partition of the corpus. For evaluation on the development set (which participants may wish to do to pre-evaluate alternative systems they develop), we optimised the parameters on the training set using 5-fold cross-validation. LibSVM for matlab was used throughout our experiments [2].

For the Word-Level Sub-Challenge (WLSC), we trained a single set of regressors using both audio and video features, as for every word there would be both audio and video features present. This is not the case for the Fully Continuous Sub-Challenge (FCSC) however. For the FCSC we consider two conditions: speech and non-speech. As the speaker state is provided with the challenge data as part of the aligned transcripts, we consider the speaker-state known for the training, development, and test partitions. We thus train a separate set of regressors for the non-speech condition using only video features, and another set of regressors for the speech condition where we concatenate the audio and video features into a single feature vector.

The memory consumption of the video features is very large: with over 1.3 million frames and 5 908 features per frame, memory capacity constraints are likely to be exceeded when training a model using all data on a desktop PC. In addition, we need to find a way to fuse the video features with the audio features. For the WLSC, a single audio feature vector is computed for every word, with words having a variable duration, while for the FCSC, a single audio feature vector is computed over every half-second segment. In contrast, video features are computed every $1/49.979 \approx 0.02$ seconds. The simplest thing to do would be to down-sample the video feature rate. But given the histogram nature of the video features, a better opportunity presents itself, and that is to take the histogram of the video features not just for a single frame, but also over multiple frames. The number of frames per segment depends on the sub-challenge: they are defined as 25 frames for the FCSC (to coincide with the audio sampling rate), and have a variable number of frames in the WLSC, defined by the word duration.

To further reduce the memory consumption of the video features, we reduced the dimensionality of the video features

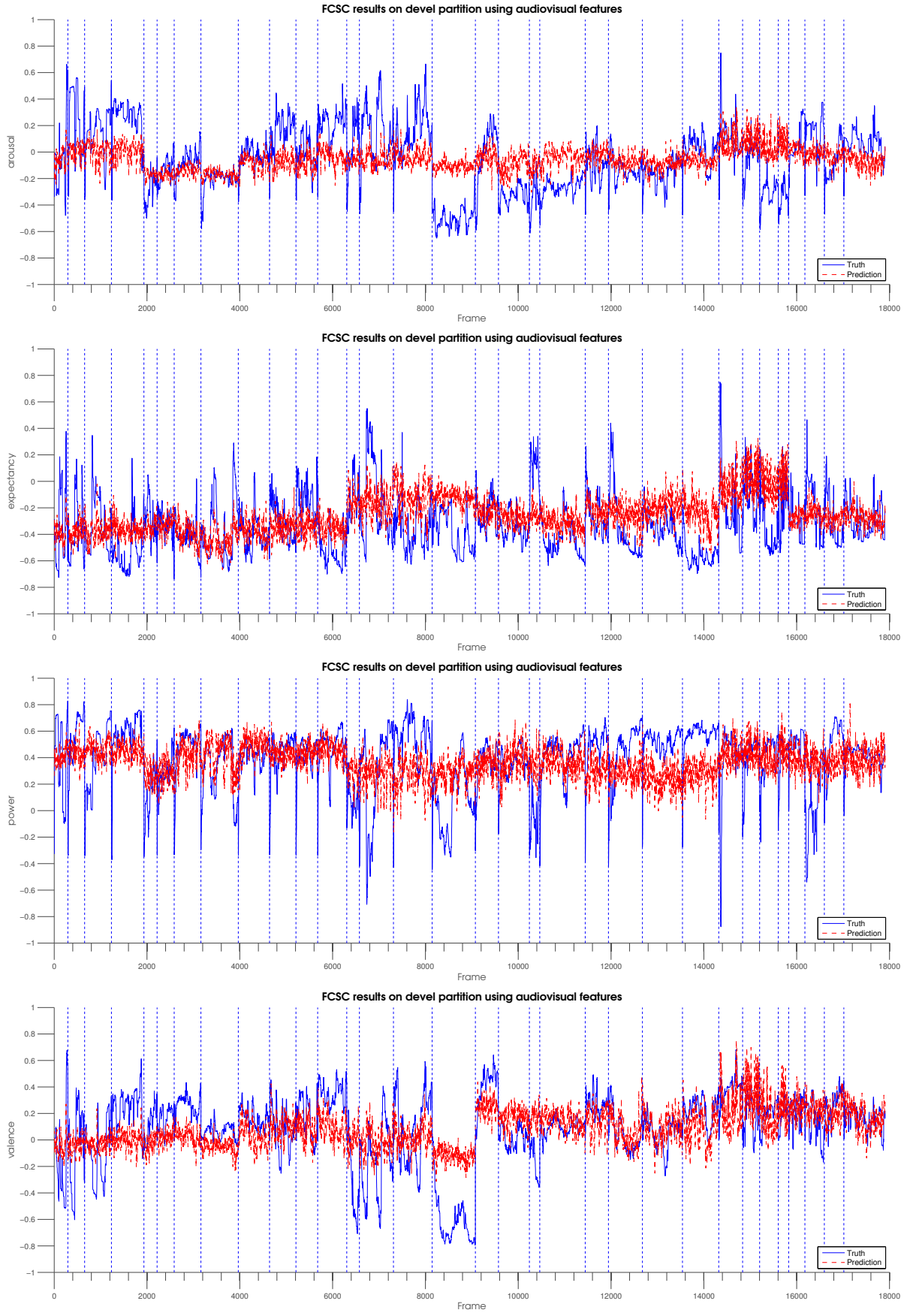

Figure 2: Baseline FCSC results on the development partition, for all four dimensions. Vertical dashed lines delimit consecutive test sessions.

Table 6: Baseline results. Performance is measured in cross-correlation averaged over all sequences.

Accuracy	Arousal	Expectation	Power	Valence	Mean
Audio-Visual					
FCSC test	0.141	0.101	0.072	0.136	0.112
WLSC test	0.021	0.028	0.009	0.004	0.015
FCSC development	0.181	0.148	0.084	0.215	0.157
WLSC development	0.018	0.009	0.001	0.002	0.007
Audio only					
WLSC test	0.014	0.038	0.016	0.040	0.027
WLSC development	0.054	0.020	0.019	0.062	0.039
Video only					
FCSC test	0.077	0.128	0.030	0.134	0.093
WLSC test	0.005	0.012	0.018	0.005	0.011
FCSC development	0.151	0.122	0.031	0.207	0.128
WLSC development	0.032	0.013	0.005	0.003	0.014

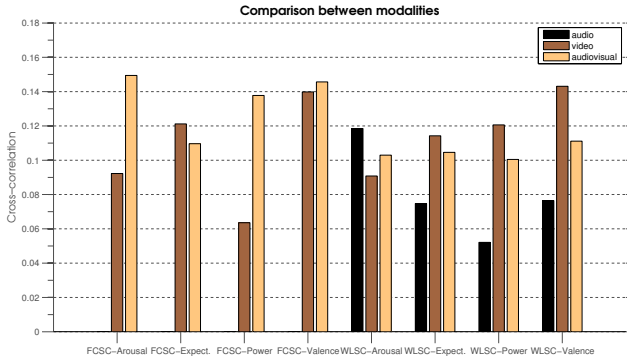

Figure 3: Comparison between the audio only, video only, and feature-level fused audio-visual modalities. Note that there are no results for the audio modality on the FCSC.

by modifying the face partition grid (see Fig 1). We removed the left-and right-most columns, as they often do'nt include the face, or at least not very relevant elements of the face. We further merged blocks in groups of 2×2, resulting in a new grid of 5 rows and 4 columns. The final dimensionality of the video feature vector is then reduced to1188.

Performance is measured in terms of correlation. To be more precise, we calculate the correlation coefficient between the predicted labels and ground truth labels per character interaction (session), per dimension, and calculate the average over all sessions and dimensions. Results for the two Sub-Challenges are given in Table 6. The top two rows are the official baseline results, that is, the FCSC and WLSC results on the test partition using audio-visual features. Participants will be ranked on the mean performance over all four dimensions, i.e. the final column in this table. The table also shows results on the development set, and results obtained on both the test and development partitions using only video or only audio features. Note that for the FCSC it isn't possible to provide audio-only results, as there are large parts of the data where the user isn't speaking. The prediction results on the FCSC development partition using Audio-Visual features is illustrated in Fig. 2.

The baseline results show that for the FCSC on the test partition (i.e. the scores that the participants will be rated on), scores lie between 3% and 14.1% correlation for the four dimensions. It also shows, that scores are consistently highest for the Valence dimension, and most often lowest for the Power dimension. The results also allow a comparison between the audio, video, and audio-visual modalities. This is further illustrated in Fig. 3. The results show that for the FCSC, fusion of audio and video modalities generally increases performance. For the WLSC however, it appears that audio is the dominant modality. This may be because the word boundaries aid forging better audio features on the one hand, and hinder creating affective video features because of mouth movements on the other hand. The fact that for the WLSC the audio and video features are temporally misaligned for a maximum of 0.5 seconds per word is another possibility for the fusion to fail.

Finally, the WLSC scores are significantly lower than the FCSC scores. This is likely due to the nature of the predicted labels, ground truth labels, and the performance measure used. The ground truth labels are a signal with a temporal resolution of 50 Hz, and can vary over the duration of a word. The predicted labels on the other hand have a variable temporal resolution (depending on word duration), and have a constant value during the entire word.

6. CONCLUSION

We introduced AVEC 2012 – the first combined open Audio/Visual Emotion Challenge to be continuous in time and value representation, its conditions, data, baseline features and results. By intention, we preferred to use open-source software and highest transparency and realism for the baselines by refraining from feature space optimisation and optimising on test data. This should improve the reproducibility of the baseline results.

With correlation coefficients on the test partition ranging between 0.03 and 0.141 for the Fully Continuous Sub-Challenge, and between 0.004 and 0.038 for the Word-Level Sub-Challenge, it is evident that continuous affect recognition from audio and/or video is indeed a daunting task. The baseline results indicate that higher scores are attained on the Fully Continuous than on the Word-Level Sub-Challenge.

Importantly, the results show that for the Fully Continuous Sub-Challenge, fusing audio and video is clearly beneficial.

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

7. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007- 2013) under grant agreement No. 289021 (ASC-Inclusion). The authors would further like to thank the sponsor of the challenge, the Social Signal Processing Network (SSPNet). The responsibility lies with the authors.

8. REFERENCES

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

[2] Chang, C.-C., and 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., and Schröder, M. Feeltrace: An instrument for recording perceived emotion in real time. In *Proc. ISCA Workshop on Speech and Emotion* (Belfast, UK, 2000), pp. 19–24.

[4] Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., and 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* (Paris, France, 2008), ELRA, pp. 1–4.

[5] Eyben, F., Wöllmer, M., and Schuller, B. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. ACII* (Amsterdam, The Netherlands, 2009), pp. 576–581.

[6] Eyben, F., Wöllmer, M., and Schuller, B. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. ACM Multimedia (MM)* (Florence, Italy, 2010), pp. 1459–1462.

[7] Eyben, F., Wöllmer, M., Valstar, M., Gunes, H., Schuller, B., and Pantic, M. String-based Audiovisual Fusion of Behavioural Events for the Assessment of Dimensional Affect. In *Proceedings International Workshop on Emotion Synthesis, rePresentation, and Analysis in Continuous spacE, EmoSPACE 2011, held in conjunction with the 9th IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, FG 2011* (Santa Barbara, CA, March 2011), IEEE, IEEE, pp. 322–329.

[8] Fontaine, J., K.R., S., Roesch, E., and Ellsworth, P. The world of emotions is not two-dimensional. *Psychological science 18*, 2 (2007), 1050 – 1057.

[9] Jiang, B., Valstar, M., and 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* (Santa Barbara, USA, 2011), pp. 314–321.

[10] Lichtenauer, J., Valstar, M. F., Shen, J., and Pantic, M. Cost-effective solution to synchornized audio-visual capture using multiple sensors. *Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance* (2009), 324–329.

[11] McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing 3* (2012), 5–17.

[12] Ojala, T., Pietikainen, M., and Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 7 (2002), 971–987.

[13] Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., de Sevin, E., Valstar, M., and Wöllmer, M. Building Autonomous Sensitive Artificial Listeners. *IEEE Transactions on Affective Computing 3* (2012), 165–183.

[14] Schuller, B., Steidl, S., and Batliner, A. The INTERSPEECH 2009 Emotion Challenge. In *Proc. INTERSPEECH 2009* (Brighton, UK, 2009), pp. 312–315.

[15] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. The INTERSPEECH 2010 Paralinguistic Challenge. In *Proc. INTERSPEECH 2010* (Makuhari, Japan, 2010), pp. 2794–2797.

[16] Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., and Pantic, M. AVEC 2011 – The First International Audio/Visual Emotion Challenge. In *Proceedings First International Audio/Visual Emotion Challenge and Workshop, AVEC 2011, held in conjunction with the International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2011, ACII 2011*, vol. II. Springer, Memphis, TN, October 2011, pp. 415–424.

[17] Shan, C., Gong, S., and Mcowan, P. W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing 27*, 6 (2009), 803–816.

[18] Valstar, M., Jiang, B., Mehu, M., Pantic, M., and Scherer, K. The first facial expression recognition and analysis challenge. *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition* (2011), 921–926.

[19] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. *The HTK book (v3.4)*. Cambridge University Press, Cambridge, UK, 2006.