String-based Audiovisual Fusion of Behavioural Events for the Assessment of Dimensional Affect

Florian Eyben, Martin Wöllmer, Michel F. Valstar, Hatice Gunes, Björn Schuller and Maja Pantic

***Abstract*— The automatic assessment of affect is mostly based on feature-level approaches, such as distances between facial points or prosodic and spectral information when it comes to audiovisual analysis. However, it is known and intuitive that behavioural events such as smiles, head shakes or laughter and sighs also bear highly relevant information regarding a subject's affective display. Accordingly, we propose a novel *string-based* prediction approach to fuse such events and to predict human affect in a continuous dimensional space. Extensive analysis and evaluation has been conducted using the newly released SEMAINE database of human-to-agent communication. For a thorough understanding of the obtained results, we provide additional benchmarks by *more conventional* feature-level modelling, and compare these and the string-based approach to fusion of signal-based features and string-based events. Our experimental results show that the proposed *string-based* approach is the best performing approach for automatic prediction of Valence and Expectation dimensions, and improves prediction performance for the other dimensions when combined with at least acoustic signal-based features.**

I. INTRODUCTION

A significant part of past research in machine analysis of human affect has focused on the recognition of prototypic expressions (i.e., of seven basic emotions) based on data that has been posed on demand and acquired in laboratory settings [1], [2]. However, it has been shown that in everyday interactions people exhibit non-basic, subtle and rather complex affective states like thinking and embarrassment [3]. Therefore, a single label (or any small number of discrete classes) may not reflect the complexity of the affective state conveyed by such rich sources of information. Hence, a number of researchers advocate the use of dimensional description of human affect, where affective states are not independent from one another; rather, they are related to one another in a systematic manner [4].

In light of these, this paper focuses on combining multiple audiovisual cues for automatic, dimensional and continuous interpretation of affective displays recorded in naturalistic settings. More specifically, we propose a novel *string-based* approach for fusing verbal (i.e., spoken words) and non-verbal behavioural events (e.g., smiles, head shakes or laughter) for automatic prediction of human affect in a continuous dimensional space. This approach stands in contrast to most conventional approaches, which are based on audio/video based feature-level modelling and fusion. As we also compute "features" for fusing the event strings, the features derived from event strings are referred to as *string-based* or *event-based features* while the (low-level) features computed directly from the audio or video signal are referred to as *signal-based features* or *signal features*.

The following subsections provide a brief introduction to the background of dimensional affect recognition and introduce related work.

A. Affect in Dimensional Space

The prosodic features which seem to be reliable indicators of the basic emotions are the continuous acoustic measures, particularly pitch-related measures (range, mean, median, and variability), intensity and duration. For a comprehensive summary of acoustic cues related to vocal expressions of basic emotions, readers are referred to [5]. There have also been a number of works that focus on how to map audio expression to dimensional models. Cowie et al. used the Valence-Activation space, which is similar to the Valence-Arousal (V-A) space, to model and assess emotions from speech [5]. Scherer and colleagues have also proposed how to judge emotion effects on vocal expression, using appraisal-based theory [6], [7].

Facial actions (e.g., pulling eyebrows up) and facial expressions (e.g., producing a smile), and to a lesser extent bodily postures (e.g., backwards head bend and arms raised forwards and upwards) and expressions (e.g., head nod), form the widely known and used visual signals for automatic affect measurement. Dimensional models are considered important in this task, as a single discrete label may not reflect the complexity of the affective state conveyed by the combination of facial expression, body posture and body gesture.

A number of researchers have investigated how to map various visual signals onto emotion dimensions. For instance, [4] mapped the facial expressions to various positions on V-A space (e.g., joy is mapped on the high arousal - positive valence quadrant), while [8] investigated the emotional and communicative significance of head nods and shakes in terms of Arousal and Valence dimensions, together with dimensional representation of Solidarity, Antagonism and Agreement.

B. Dimensional Affect Recognition from Audio and Video

Automatic dimensional affect recognition is still in its pioneering stage [1], [9],[10],[11],[12]. The most commonly employed strategy is to reduce the dimensional emotion

Florian Eyben, Martin Wöllmer, and Björn Schuller are with the Institute for Human-Machine Communication, Technische Universität München. At the time of writing, Björn Schuller was a visiting researcher in the Department of Computing, Imperial College London.

Michel Valstar, Hatice Gunes, and Maja Pantic are with the iBUG group, Department of Computing, Imperial College London. Maja Pantic is also with EEMCS, Twente University.

classification problem to a two-class problem (positive vs. negative or active vs. passive classification; e.g., [13],[14]) or a four-class problem (classification into the quadrants of 2D V-A space; e.g., [15], [16], [17], [18], [19]).

In dimensional affect recognition emotions are represented along a continuum. Considering this, most systems that target automatic dimensional affect recognition tend to simplify the problem by quantising the continuous labels into a finite number of discrete levels. For example, Kleinsmith and Bianchi-Berthouze discriminate between high-low, high-neutral and low-neutral affective dimensions [20], while Wöllmer et al. quantise the V-A dimensions of the SAL database into either 4 or 7 levels, and then use Conditional Random Fields (CRFs) to predict the quantised labels [10] . Attempts for discriminating between more coarse categories, such as positive vs. negative [13], and active vs. passive [15] have also been attempted. Of these, Caridakis et al. [15] uses the SAL database, combining auditive and visual modalities. Nicolaou et al. focus on audio-visual classification of spontaneous affect into negative or positive emotion categories using facial expression, shoulder and audio cues, and utilising 2- and 3-chain coupled Hidden Markov Models and likelihood space classification to fuse multiple cues and modalities [13]. Kanluan et al. combine audio and visual cues for affect recognition in V-A space by fusing facial expression and audio cues, using SVRs and late fusion with a weighted linear combination [21] with discretised labels (on a 5-point scale in the range of [-1,+1] for each emotion dimension). The work presented in [19] utilises a hierarchical dynamic Bayesian network combined with BLSTM-NN performing regression and quantising the results into four quadrants (after training).

As far as actual continuous dimensional affect prediction (without quantisation) is concerned, four attempts have been proposed so far, two of which deal exclusively with speech (i.e., [10], [22]). The work by Wöllmer et al. uses Long Short-Term Memory neural networks and Support Vector Machines for Regression (SVR) [10]. Grimm and Kroschel use SVRs and compare their performance to that of the distance-based fuzzy k-Nearest Neighbour and rule-based fuzzy-logic estimators [22]. The work by Gunes and Pantic focuses on dimensional prediction of emotions from spontaneous conversational head gestures by mapping the head motion vectors and occurrences of head nods and shakes into Arousal, Expectation, Intensity, Power and Valence level of the observed subject using SVRs [23]. The work by Nicolaou et al. focuses on dimensional and continuous prediction of emotions from naturalistic facial expressions within the context of an Output-Associative Relevance Vector Machine regression framework that augments the traditional Relevance Vector Machine regression by learning non-linear input and output dependencies inherent in the affective data [24].

For further details on the aforementioned systems, as well as on systems that deal with dimensional affect recognition from a single modality or cue, the reader is referred to [1], [2], [12].

In summary, none of the related works have investigated *string-based* prediction and multimodal fusion of verbal and nonverbal behavioural events for automatic prediction of human affect in a continuous dimensional space.

The remainder of this paper is structured as follows: In Section II the corpus used for the experimental validation, i.e., the SEMAINE database of human-agent communication, is shortly introduced. We describe the methods used for automatic behavioural event detection and classification by video and audio analysis in Section III. The experimental setup and the string-based multimodal fusion of the behavioural events can be found in Section IV and Section V, respectively. For comparison, we then introduce a more conventional fusion approach to audiovisual affect analysis in Section VI, before discussing the results in Section VII and drawing our conclusions in Section VIII.

II. THE SEMAINE DATABASE

The SEMAINE database [25] was recorded to study natural social signals that occur in conversations between humans and the future generation of artificially intelligent agents, and to collect data for the training of such intelligent agents. The scenario used for this is called the Sensitive Artificial Listener, SAL for short. 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 he/she says. The model is a style of interaction observed in chat shows and parties, which aroused interest because it seems possible that a machine with some basic emotional and conversational competence could sustain such a conversation, without needing to be competent with fluent speech and language understanding.

In the recording scenario, the participants are asked to talk 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 depressive.

The study presented in this work is based on the first part of the SEMAINE database. In this part, human operators pretended to be the artificial agents. This type of interaction is called Solid-SAL. Because we assume that the SAL agent has no language understanding, a few rules govern this type of interaction. The most important of these is that the agent is not allowed to answer questions. However, the operators are instructed that the most important aspect of their task is to create a natural style of conversation; strict adherence to the rules of a SAL engagement was secondary to a conversational style that would produce a rich set of conversation-related behaviours and therefore transgressions occasionally occur.

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. Both the user and the operator are recorded frontally by both a greyscale camera and a colour camera. In addition, the user is recorded by a greyscale camera positioned on one side of the user to capture a profile view of their face and body. 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. [26].

The Solid-SAL part of the database holds recordings of 20 trials of the SAL experiment, split into over 100 character conversations of approximately 5 minutes each. All recorded conversations have been fully transcribed and annotated for five affective dimensions and partially annotated for 27 other dimensions, using trace style continuous ratings. The five core dimensions are those that psychological evidence suggests are best suited to capture affective colouring in general [27]. They are *Valence*, *Activation*, *Power*, Anticipation/*Expectation* and (overall emotional) *Intensity*.

Further details on the SEMAINE database can be found in [25]. The database is freely available for scientific research purposes from http://semaine-db.eu.

III. BEHAVIOURAL EVENTS

This section describes the procedures employed to detect the behavioural events that are used for the proposed string-based affect prediction and fusion approach.

A. Nonverbal Visual Events

The nonverbal events detected from the visual modality are head gestures and facial action units (AU). Once detected, these events are supplied as features to the string-based prediction and fusion algorithm. Due to lack of annotated SEMAINE data (in terms of head gestures and AUs), how each visual event detection component affects the string-based prediction algorithm and its accuracy could not be evaluated.

Head gestures. We aim to recognise four different head gestures: head nods, head shakes, head tilts to the left, and head tilts to the right. The automatic detection of head nods and shakes is based on the 2-dimensional (2D) global head motion estimation. The face region is detected using the well known Viola and Jones face detector [28]. In order to determine the magnitude and the direction of the 2D head motion, optical flow is computed between two consecutive frames. It is applied to a refined region (i.e., resized and smoothed) within the detected facial area to exclude irrelevant background information.

After preliminary analysis, the angle component of the 2D head motion vector has been considered as the distinguishing feature in order to represent nods and shakes. The angle measure has then been discretised by representing it with directional codewords. The directional codeword is obtained by quantising the direction into four codes for head movements (for rightward, upward, leftward and downward motion, respectively) and one for 'no movement'. The directional codewords generated by the visual feature extraction module are then fed into a Hidden Markov Model (HMM) for training a *nodHMM* and a *shakeHMM*. However, to be able to distinguish other head movements from the actual head nods/shakes, we (i) threshold the magnitude of the head motion, (ii) build an *otherHMM* to be able to recognise any head movement that are not nods/shakes, and (iii) statistically analyse the likelihoods outputted by the nod/shake/other HMM (maximum likelihood vs. training classifiers on the outputted likelihoods).

TABLE I

Comparative results obtained with respect to (i) thresholding the normalised head motion magnitude, (ii) deciding on the number of states to be used within the HMM models, and (iii) whether to use likelihood space classification or maximum likelihood classification.

threshold for normalised head motion magnitude	**number of states used in the 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

152 head *nod*, 103 head *shake*, and 140 *other* clips (of variable length) were manually extracted from the SEMAINE database to train the HMM models. In order to determine how to make the final decision, evaluation has been carried out (using the aforementioned data and adopting 10-fold cross-validation) with the following criteria: (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, similarly to [13]). Table I presents the best results. 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, we opted for likelihood space classification, setting the threshold=25, and number of states=2.

In order to analyse the visual data continuously, we empirically chose a window size of 0.4 seconds (about 20 video frames) that allows the detection of both brief and longer instances of head nods/shakes (similarly to other related work). From the global head motion features extracted and the head movements (nod or shake) detected, we created a window-based feature set presented in Table II. The ground-truth for the window at hand consists of the dimensional annotations averaged over that window, for each coder separately. Please see [23] for details.

The spotting capability of the automatic head nod and shake detector was evaluated using a subset of the SEMAINE database. There exists no publicly available (audio-)visual data set annotated for head nods and shakes, at either frame-level (frame-by-frame) or event-level (where a nod starts and ends). Therefore, one of the authors manually annotated a subset of the SEMAINE database that consisted of data from 4 subjects (2 male and 2 female), over 7 sessions, and 44,060 video frames in total. As the focus of this paper is

TABLE II

Head features extracted within a fixed window of 0.4 s.

Features (16) & their description
duration of no movement
duration of the upward head movement
duration of the downward head movement
duration of the rightward head movement
duration of the leftward head movement
average of the magnitude values
variance of the magnitude values
average of the angle values
variance of the angle values
loglikelihood outputted by nodHMM
loglikelihood outputted by shakeHMM
loglikelihood outputted by otherHMM
result of the maximum likelihood classification
result of the likelihood space classification (nod vs. shake)
result of the likelihood space classification (nod vs. other)
result of the likelihood space classification (shake vs. other)

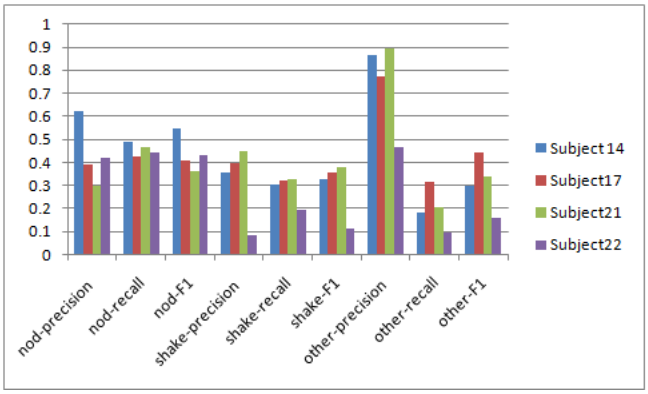

Fig. 1. Nod, shake and other event detection results (per subject) on a subset of the SEMAINE database.

on events, an event-level evaluation was conducted based on a window of 20 frames (used for decision making by the detector) by taking the majority vote as the label of the window at hand. The results are presented in Fig. 1. The figure shows that *nod event* detection seems to be best for subject 14, followed by subject 22; *other event* detection seems to be best for subject 21 followed by subject 14. *Shake event* detection appears to be best for subject 21, followed by subject 17. This in turn implies that naturalistic emotional displays are rather subject-specific in nature. However, it is difficult to draw hard conclusions given the limited amount of data. Within the SEMAINE database, the amount of nod, shake and other events varies between recording sessions and between subjects. For instance, the aforementioned test set contains 33,328 frames of other event, 6,873 frames of nod event, and 3,859 frames of shake event.

To detect head tilts, we employ a haar-cascade eye detector. The detector usually returns multiple detections per eye. To select which one is the real location of the eye, we modify the probability of each candidate location in two ways. Firstly, the probability of a candidate location is decreased according to a Gaussian function of the distance to the prior probability of the location of an eye given the detected face location. Secondly, we modify the probability of each candidate by the distance to other candidates. Candidates that are close together will increase each other's probability. This results in the predicted locations of the left-eye $\{x_l, y_l\}$ and right-eye $\{x_r, y_r\}$.

Using the locations of the centres of the eyes, we can now compute the roll of the face as $\alpha = \arctan(y_r - y_l)/(x_r - x_l)$, which, in turn, indicates whether a head tilt has occurred. Similarly to the nod/shake detection, we average α over a time window of 0.4 seconds. If the average value is greater than 0.1 radians, we say that a right-head-tilt occurred, and if it is smaller than -0.1 radians, a left-head-tilt is detected.

Facial Action Units. To detect facial Action Units (AUs), we employed the method proposed by Jiang et al. [29]. In their work the authors investigate the possibility to detect AUs using two static and two dynamic appearance descriptors. From those four we chose to use the Local Binary Patterns (LBP) descriptor. Although according to their reports the LBP descriptor did not attain the highest recognition performance, it was by far the fastest. Since the data we process in this study consist of over a million frames, speed was of great importance.

The LBP descriptor is computed by systematically comparing the central pixel with a number of surrounding pixels in a local neighbourhood. If the surrounding pixel has a higher intensity than the central pixel, the result is a binary 1, otherwise it is a 0. The results of all neighbours together forms a binary word, which is translated to a decimal number. In our case, we use the 8 immediately surrounding pixels, and thus we have an 8-bit word, and the decimal number lie in the range [0, 255]. The LBP operator is applied to all pixels in an image, and a histogram of the LBP output per pixel is created which describes the texture of that image.

To encode local texture instead of a single texture for the entire face, we divide the face region into 10 x 10 blocks. An LBP histogram is calculated for each of those blocks separately, after which the histograms of all blocks are concatenated to form a single feature vector. GentleBoost feature selection is applied to this, and the reduced feature set is fed to a bank of Support Vector Machine classifiers, one for every AU detected. Currently, the system can reliably detect 12 AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU12, AU14, AU15, AU20, AU25, and AU45). To be able to deal with appearance variation due to head roll and different sizes of faces, we use the locations of the eyes found during head tilt detection. The input images are first rotated by α radians, and then scaled to make the distance between the centres of the eyes equal to 80 pixels.

Because it is notoriously time-consuming to create ground-truth labelling of AUs from video, there is currently very little AU annotation available for the SEMAINE database. To wit, at the time of writing 181 frames have been annotated, taken from eight character conversations of two subjects, i.e., for both subjects the conversations with all four SAL characters were used. Besides testing on the SEMAINE database, we therefore also test our AU detector on 1504

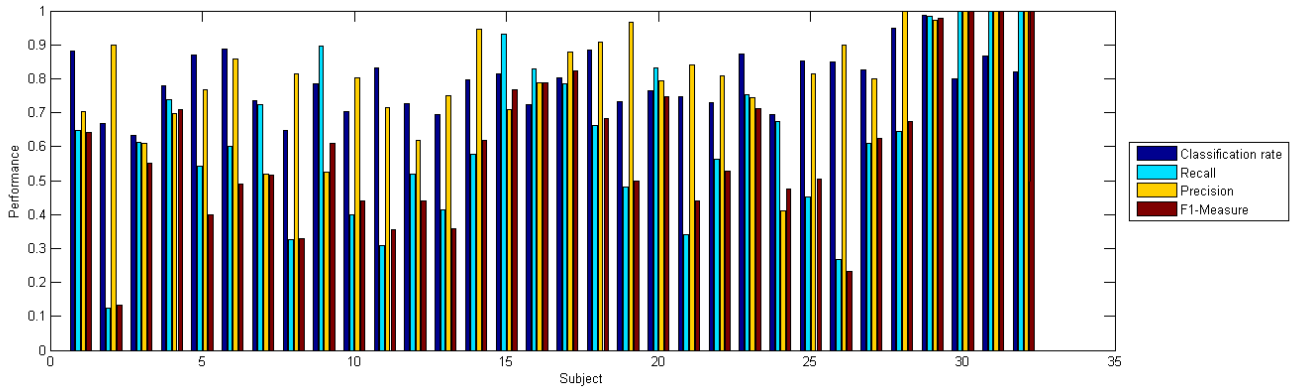

Fig. 2. AU detection result per subject on the MMI-Facial Expression and SEMAINE databases.

images of posed facial expressions taken from the MMI Facial Expression Database [30]. Tests were done in a subject-independent manner. Fig. 2 shows the average F1-measure for all AUs per subject. Subjects two and three in the figure are the two subjects taken from the SEMAINE database. It is hard to draw conclusions for the SEMAINE data given the limited data, and indeed the figure shows that although performance is competitive for Subject three, it is rather poor for Subject two. Unfortunately, there exist no freely available databases of spontaneous data with AU labelling, therefore, we cannot compare our results with those of others. The results for the subjects from the MMI Facial Expression database are all competitive with the current state of the art.

B. Verbal and nonverbal acoustic events

As acoustic events we used laughter and sighs, as they occur frequently in spontaneous emotional speech and carry substantial emotional meaning. We additionally used emotionally relevant keywords which are derived per dimension by feature selection from automatic ASR transcriptions of the whole SEMAINE database. We decided not to use the ground truth transcriptions, but the recogniser's output – also for extracting string-based features on the training set – to avoid a mismatch between ideal training conditions and imperfect recognition conditions in a real-world system.

For keyword detection we used a multi-stream large vocabulary continuous speech recognition (LVCSR) engine tuned for robust recognition of spontaneous and emotional speech (for details see [31]). In addition to the standard set of 39 cepstral mean normalised MFCC features, the system uses discrete phoneme prediction features generated by a Long Short-Term Memory (LSTM) network. The LSTM principle enables long-range context modelling on the feature level and was shown to be well-suited for modelling conversational speech [32]. Instead of conventional hidden units which can be found in the hidden layer of standard recurrent neural networks, an LSTM network consists of recurrently connected memory blocks that can store information over long time periods and are able to model co-articulation effects in human speech.

Combined with context-sensitive triphone Hidden Markov Models, the system achieved a true positive rate of 76.58 % at a false positive rate of 0.94 % when trained on the SEMAINE database, the SAL corpus, and on the COSINE corpus [33] consisting of conversational, disfluent, and partly noisy speech. The multi-stream LVCSR engine uses the on-line LSTM decoder integrated in the open-source speech feature extractor openSMILE [34] as well as a trigram language model trained on the aforementioned speech corpora. All phoneme HMM consist of three emitting states with each state having 16 Gaussian mixtures. The nonverbal events *laughing* and *sighing* are detected within the same recogniser framework. We trained HMM comprising nine hidden states for these vocalisations. The LSTM network for phoneme prediction is composed of 128 memory blocks and the size of the used vocabulary is 7.0 k.

From the 7.0 k words in the vocabulary, we selected a subset of words relevant for each of the five affect dimensions using the Correlation based Feature Subset Selection (CFS) algorithm.

IV. EXPERIMENTAL SETUP

For the experiments reported in this paper we train Support Vector Regressors (epsilon SVR with a polynomial (linear) kernel), since SVR is known to handle large feature spaces reliably. The trained models are evaluated using the SEMAINE database, using sessions that have been coded by the same three raters. Recordings 4, 6, 9, 10, 11, 13, 15, 16, 17, 18, and 19 were used for training, and recording 3, 5, 12, and 14 for testing.

As metrics for evaluation, the Mean Linear Error (MLE) and correlation coefficient (CC) are used. MLE measures the average of the absolute error between an estimator and the true value of the quantity being estimated. CC (usually referred to as Pearson's correlation) indicates the strength of a linear relationship between two variables. MLE and CC have been calculated both for individual raters and the (automatic) predictor. Both MLE and correlation have been calculated for each rater with respect to other raters and by

averaging the obtained results.

For the audiovisual analysis conducted within this paper, we only consider regions where the subjects are talking, i.e., user speech turns. Since these turns themselves are unsuitable as units of analysis due to their high variability in length (from few seconds up to minutes), we decided for an incremental segmentation scheme. This scheme has been developed for the real-time SEMAINE demonstrator system, where low-latency incremental estimation of the user's affective state is required. The turns are split into overlapping segments, which are not longer than five seconds and are sampled every second. Thus, the first segment within a turn spans the range from 0 s to 1 s, the second from 0 s to 2 s, the fifth from 0 s to 5 s, and the sixth segment from 1 s to 6 s, and so on. A continuous affect label for each dimension is assigned to each segment by simple averaging of the dimensional affect labels within the segment. Applying the aforementioned segmentation procedure leads to 7,699 segments in the training set, and 1,324 segments in the evaluation set.

V. STRING-BASED FUSION

The event fusion is performed at the string-level per segment (see section II for a definition) by joining all events where more than half of the event overlaps with the segment in a single string. The events can thus be seen as "words". The resulting strings are converted to a feature vector representation through a binary bag-of-words (BOW) approach. By doing so we do not consider term frequencies, i.e., we only consider whether a certain event is present or not within a segment and do not count how often events occur. We decided to use this simple approach because, in contrast to the keywords and vocal outbursts, the video-based events are not identified as unique instant events in time, but only locally as predictions for short time frames. Some post-processing would have to be applied in order to group these predictions into discrete events, which we will carefully attempt to do as the next step in future work.

Due to the large vocabulary size in the corpus, we have to select emotionally relevant words from the approximately 7.0 k dimensional word vector. We do this separately for each of the five dimensions using CFS as a feature selection algorithm. Approximately 200–300 words remain after this feature selection. We add laughter and sigh BOW features to the reduced word vector to obtain the audio event vector (Event A). The video event vector (Event V) contains two BOW dimensions for nod/shake, 12 dimensions for AUs, and two dimensions for tilt left/right. We do not apply feature selection here, thus this vector is always 16 dimensional.

The results of the string-based emotion recognition are given in table III (rows labelled with Event A/V). Results for conventional acoustic and video signal-based feature approaches are also provided for comparison, as well as results for fusion of events with signal-based features. The signal-based features are described in the next section.

At this point we would like to point out that all the event-based features used in this paper have been computed on the actual output of the event detectors and *not* on the ground truth labels, i.e. we are presenting fully realistic processing conditions.

TABLE III

All results for affect prediction for five continuous dimensions A(ctivation), V(alence), E(xpectation), I(ntensity), P(ower). Target label is the mean of Rater 3, 5, and 6 annotations. SVR regression with polynomial kernel of degree 1. Correlation coefficient (CC) and Mean Linear Error (MLE). Audio (A): audio features (functionals of acoustic LLD); Video (V): functionals of 2D head motion-based features (nod/shake); Event A/V: String-based features from audio (A) events (words and laughs/sighs) and/or video events (action units and head nod/shake/tilt). Best result(s) printed in bold face.

CC	**A**	**V**	**E**	**I**	**P**
Audio (A)	0.653	-0.085	0.190	0.503	0.367
Video (V)	0.204	0.037	0.037	0.397	-0.019
Event A+V	0.447	**0.165**	0.220	0.397	0.264
Event A	0.215	0.123	**0.282**	0.148	0.275
Event V	0.524	-0.014	-0.254	0.421	-0.013
A + V + Event A+V	**0.699**	0.037	0.213	**0.548**	0.405
A + V	0.661	-0.103	0.191	0.573	0.338
A + Event A+V	**0.699**	0.092	0.218	0.525	**0.431**
MLE	**A**	**V**	**E**	**I**	**P**
Audio (A)	0.157	0.265	0.181	0.195	0.173
Video (V)	0.208	0.258	0.185	0.194	0.183
Event A+V	0.188	0.255	0.180	0.199	0.181
Event A	0.206	**0.245**	**0.173**	0.211	0.177
Event V	0.187	0.271	0.194	0.204	0.188
A + V + Event A+V	**0.153**	0.271	0.180	**0.183**	0.171
A + V	0.156	0.282	0.180	0.185	0.175
A + Event A+V	0.154	0.259	0.181	0.189	**0.170**

VI. FEATURE-LEVEL FUSION AND COMPARATIVE ANALYSIS

This section aims to provide a baseline for comparing the newly introduced string-based prediction and fusion, and the traditional signal-based approaches and feature-level fusion. In addition to these, fusion of string-based features with signal-level features is also employed for further analysis.

The signal-level audio feature set is based on the one used for the baseline results of the INTERSPPECH 2010 Paralinguistic Challenge [35]. This 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 20–6,500 Hz). In order to improve the computational efficiency for real-time on-line processing in the SEMAINE demonstrator system, we decided to omit the line spectral pairs as low-level features and remove the percentile functionals (quartiles, and inter-quartile ranges), which require the low-level feature contours to be sorted with quick-sort. In total this leads to a 1,880 dimensional feature set: 47 low-level descriptors, first order delta coefficients, and 20 functionals yields 1,880 features. Including the number of pitch onsets and the total segment duration in seconds gives the final number of 1,882. A description of the feature set is given in table IV.

The extracted video features related to head gestures are presented in Table II). After the feature extraction, the 20

TABLE IV

Acoustic features.

Descriptors (47)	**Functionals (20)**
Loudness, Intensity, RMS & LOG energy	min., max. value and range
Voicing Probability	rel. position of max / min value
F0 (pitch) only in voiced regions	arithmetic mean
MFCC 0–12	slope, offset, lin. and quad. error
RASTA style PLP-CC 0–7	standard deviation, skewness, kurtosis
MFB 1–14	time signal is above 25 %, 50 %, 75 %, and 90 %
Spectral Flux, Centroid, Entropy, Variance	time signal is below 50 %
95% spectral roll-off point	
Mean crossing rate (time-domain)	

functionals listed in table IV are applied to these features. Thus, a single vector of video features is created for each segment, which can easily be concatenated with the acoustic feature vector and the string-based bag-of-words vector.

VII. DISCUSSION OF RESULTS

The results – as shown in table III – clearly show that the proposed string-based approach for multimodal affect prediction is feasible and gives the best result for the dimensions Valence and Expectation. This is in line with findings that these dimensions are poorly correlated with acoustic features alone, for example. The approach also improves the predictors' performance if combined with signal-based features. The overall best result is achieved for Activation, where the average result is as good as human performance.

TABLE V

Correalation coefficient (CC) and Mean Linear Error (MLE) for five affect dimensions A(ctivation), V(alence), E(xpectation), I(ntensity), and P(ower) of the three human coders computed for each coder as the MLE or CC between the coder's annotation and the mean of the other two coders' annotations on the test set.

CC	**A**	**V**	**E**	**I**	**P**
R3	0.748	0.835	0.462	0.788	0.487
R5	0.757	0.776	0.418	0.763	0.483
R6	0.607	0.844	0.261	0.688	0.143
mean	**0.704**	**0.818**	**0.380**	**0.746**	**0.371**
MLE	**A**	**V**	**E**	**I**	**P**
R3	0.429	0.159	0.262	0.322	0.309
R5	0.367	0.174	0.434	0.252	0.241
R6	0.199	0.152	0.340	0.191	0.346
mean	**0.332**	**0.162**	**0.345**	**0.255**	**0.299**

Table V gives the performance of each human annotator compared to the average of the other two annotators. We can see that the performance of our automatic predictors is not yet at the level of human performance for all five dimensions, but we are getting quite close for some dimensions, Activation and Power dimensions, in particular. A huge difference still remains for the Valence dimension, where human performance/agreement is highest among all five dimensions, but the correlation of the automatic prediction is lowest. Considering the fact that the *Event A+V* and *Event A* features gave best and second best results for automatic prediction of Valence, this could be seen as an indication that annotators strongly take the content and meaning of utterances into account when creating their judgements. Another notable issue that is evident when comparing human and automatic performance is that the MLE is much lower for automatic prediction than for human coder agreement (except for Valence). This can be attributed to the fact that human annotators use their individual scalings and offsets when performing the annotations, which results in a higher error but does not affect the overall correlation. Automatic predictors generally try to optimise the output error during training. Thus, for future continuous dimensional affect prediction systems we should focus on the correlation coefficient as a main evaluation metric, as followed in the INTERSPEECH 2010 Paralinguistic Challenge [35].

VIII. CONCLUSION AND OUTLOOK

We have investigated a novel approach to audiovisual fusion on the SEMAINE database. The approach is based on the bag-of-words technique which is already well known and used for linguistic emotion recognition. We extended this approach to multimodal string-based fusion by adding video-based events (facial expression Action Units, head nods, shakes, and tilts) as 'words' to the string of acoustic events. We have also compared the proposed approach to traditional signal-feature-based approaches and have investigated the potential of fusing features from the proposed string-based approach and signal-based features (audio and video), which gave the best performance for three out of five affect dimensions.

Future work will investigate novel feature types as well as further combinations of feature groups and modalities to improve the prediction performance, especially for the Valence dimension. We will also investigate scaling and offset correction as well as smoothing for the individual annotator tracks of the SEMAINE database as a pre-processing step in order to obtain a more universal and noise free ground truth.

In the light of our results we can conclude that the proposed string-based approach is the best performing approach for automatic prediction of Valence and Expectation dimensions, and improves prediction performance for the other three dimensions, when combined with signal-based features. For Activation a correlation coefficient of 0.70 and for Power of 0.43 is obtained in this case. This is as good or even slightly better than human performance.

IX. ACKNOWLEDGMENTS

This work has been funded by the European Community's 7th Framework Programme [FP7/2007-2013] under grant agreement no 211486 (SEMAINE). The work of M. Valstar and M. Pantic is also funded in part by the European Communitys 7th Framework Programme [FP7/20072013] under the grant agreement no 231287 (SSPNet).

REFERENCES

[1] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

[2] Z. Zeng, M. Pantic, Glenn I. Roisman, and T.S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[3] S. Baron-Cohen and T. H. E. Tead, *Mind reading: The interactive guide to emotion*, Jessica Kingsley Publishers Ltd., 2003.

[4] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human–computer interaction," *IEEE Signal Processing Magazine*, vol. 18, pp. 33–80, 2001.

[6] K.R. Scherer, *The Neuropsychology of Emotion*, chapter Psychological models of emotion, pp. 137–162, Oxford University Press, 2000.

[7] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and Cognition*, vol. 17, pp. 484–495, 2008.

[8] R. Cowie, H. Gunes, G. McKeown, L. Vaclau-Schneider, J. Armstrong, and E. Douglas-Cowie, "The emotional and communicative significance of head nods and shakes in a naturalistic database," in *In Proc. of LREC Int. Workshop on Emotion*, 2010, pp. 42–46.

[9] M. Grimm, E. Mower, K. Kroschel, and S. Narayanan, "Primitives based estimation and evaluation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[10] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. Interspeech*, Brisbane, 2008, pp. 597–600.

[11] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," *Image and Vision Computing Journal*, vol. 27, pp. 1760–1774, 2009.

[12] H. Gunes and M. Pantic, "Automatic measurement of affect in dimensional and continuous spaces: Why, what, and how?," in *Proc. of Measuring Behavior*, 2010, pp. 122–126.

[13] M.A. Nicolaou, H. Gunes, and M. Pantic, "Audio-visual classification and fusion of spontaneous affective data in likelihood space," in *Proc. of IEEE Int. Conf. on Pattern Recognition*, 2010, pp. 3695–3699.

[14] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Gerhard Rigoll, and Andreas Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proceedings of ASRU*. 2009, IEEE.

[15] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis, "Modeling naturalistic affective states via facial and vocal expressions recognition," in *Proc. of ACM Int. Conf. on Multimodal Interfaces*, 2006, pp. 146–154.

[16] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[17] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," in *Proc. of Computer Vision and Pattern Recognition Workshops*, 2008, pp. 1–6.

[18] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, and S. Kollias, "Emotion recognition through facial expression analysis based on a neurofuzzy method," *Journal of Neural Networks*, vol. 18, pp. 423–435, 2005.

[19] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867 –881, Oct. 2010.

[20] A. Kleinsmith and N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," in *Proc. of the Int. Conf. on Affective Computing and Intelligent Interaction*, 2007, pp. 48–58.

[21] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-visual emotion recognition using an emotion recognition space concept," *Proc of the 16th European Signal Processing Conf.*, 2008.

[22] M. Grimm and K. Kroschel, "Emotion estimation in speech using a 3d emotion space concept," in *In Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2005, pp. 381–385.

[23] H. Gunes and M. Pantic, "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners," in *Proc. of Int. Conf. on Intelligent Virtual Agents*, 2010, pp. 371–377.

[24] M.A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," in *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2011.

[25] G McKeown, M.F Valstar, M Pantic, and R Cowie, "The semaine corpus of emotionally coloured character interactions," *Proc. Int'l Conf. Multimedia & Expo*, pp. 1–6, Jan 2010.

[26] J.F Lichtenauer, J Shen, M.F Valstar, and M Pantic, "Cost-effective solution to synchronised audio-visual data capture using multiple sensors," *Journal of Visual Communication and Image Representation*, pp. 1–39, Nov 2010.

[27] J.R.J Fontaine, Scherer K.R., E.B Roesch, and P.C Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 2, pp. 1050 – 1057, Feb 2007.

[28] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. 511–518.

[29] B. Jiang, M.F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. IEEE Int'l conf. on Face and Gesture recognition*, Mar 2011.

[30] M.F Valstar and M Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," *Proceedings of the Language Resources and Evaluation Conf.*, pp. 65–70, Mar 2010.

[31] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "Recognition of spontaneous conversational speech using long short-term memory phoneme predictions," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 1946–1949.

[32] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework," *Cognitive Computation*, vol. 2, no. 3, pp. 180–190, 2010.

[33] A. Stupakov, E. Hanusa, J. Bilmes, and D. Fox, "COSINE - a corpus of multi-party conversational speech in noisy environments," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[34] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia*, Firenze, Italy, 2010.

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