Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction[1]

Kevin Brady
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420, USA
781-981-6926
kbrady@LL.MIT.EDU

Youngjune Gwon
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420, USA
781-981-6365
gyj@LL.MIT.EDU

Pooya Khorrami
Beckman Institute
University of Illinois
405 N. Matthews Ave.
Urbana, Illinois 61801, USA
pkhorra2@ILLINOIS.EDU

Elizabeth Godoy
MIT Lincoln Laboratory
781-981-8287
elizabeth.godoy@LL.MIT.EDU

William Campbell
MIT Lincoln Laboratory
781-981-1751
wcampbell@LL.MIT.EDU

Charlie Dagli
MIT Lincoln Laboratory
781-981-4430
dagli@LL.MIT.EDU

Thomas S Huang
Beckman Institute
University of Illinois
t-huang1@ILLINOIS.EDU

ABSTRACT
The automatic determination of emotional state from multimedia content is an inherently challenging problem with a broad range of applications including biomedical diagnostics, multimedia retrieval, and human computer interfaces. The Audio Video Emotion Challenge (AVEC) 2016 provides a well-defined framework for developing and rigorously evaluating innovative approaches for estimating the arousal and valence states of emotion as a function of time. It presents the opportunity for investigating multimodal solutions that include audio, video, and physiological sensor signals. This paper provides an overview of our AVEC Emotion Challenge system, which uses multi-feature learning and fusion across all available modalities. It includes a number of technical contributions, including the development of novel high- and low-level features for modeling emotion in the audio, video, and physiological channels. Low-level features include modeling arousal in audio with minimal prosodic-based descriptors. High-level features are derived from supervised and unsupervised machine learning approaches based on sparse coding and deep learning. Finally, a state space estimation approach is applied for score fusion that demonstrates the importance of exploiting the time-series nature of the arousal and valence states. The resulting system outperforms the baseline systems [10] on the test evaluation set with an achieved Concordant Correlation Coefficient (CCC) for arousal of 0.770 vs 0.702 (baseline) and for valence of 0.687 vs 0.638. Future work will focus on exploiting the time-varying nature of individual channels in the multi-modal framework.

AVEC'16, October 16 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2988257.2988264

General Terms
Algorithms, Experimentation

Keywords
Affective Computing, Emotion Recognition, Speech, Deep Learning, CNN, Sparse Coding, Facial Expression, Challenge

1. INTRODUCTION
The 2016 Audio-Visual Emotion Challenge (AVEC 2016) [10] aims to compare multimedia processing and machine learning methods for automatic speech, video, and physiological analysis of human emotion measured in arousal and valence.

Our audio channel approach is first based on acoustic analysis of a subject's speech utterance. In addition to the precomputed acoustic features that come with the dataset, we use our speech tools to extract acoustic features such as auditory loudness, pitch variation, and speaking rate along spectral tilt captured in the low cepstral coefficients. Moreover, we apply sparse coding, an unsupervised learning method, on the extracted audio speech features to compute the input vectors for our regressors based on support vector machine (SVM) and softmax regression neural network. We find that these features computed on the acoustic analysis are particularly superior in arousal prediction.

Both SVM and (recurrent) neural network based regression have been known for their robustness in emotion prediction tasks [16, 17]. Despite its simplicity, linear SVM (or SVM regression)

[1] This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

is proven effective for the past AVEC challenges and chosen as the baseline method [10, 18], yielding on par with or better prediction performance with many other state of the art machine learning approaches. Mel-Frequency Cepstral Coefficient (MFCC) and Shifted Delta Cepstrum (SDC) features are popular for many language and speaker identification tasks [20] when coupled with higher-level feature learning frameworks such as Gaussian mixture model (GMM) [21]. For our case, we employ sparse coding as a higher-level learning mechanism for SDC to discover useful representations for regressing the emotion dimensions by spectro-temporal decomposition of speech signals. Sparse coding has been known for state-of-the-art performances in discriminative computer vision and object recognition tasks [22, 23, 24].

For our video channel processing our approach primarily uses the deep neural network models from Khorrami et al. [6] to predict the arousal and valence scores from the video data. Previous methods [2, 4, 9] have shown the benefits of using recurrent neural networks (RNNs) to improve performance, however their methods were trained on hand-crafted features (e.g., LGBP-TOP). Recent evidence from various areas of computer vision including emotion recognition [5, 7] has shown how learned feature representations like convolutional neural networks (CNNs) can achieve superior performance to hand-crafted features. Despite these findings, few works [1, 3] have considered the merit of passing CNN features as input to the RNN.Our model first trains a single frame CNN to predict the output label. The pre-trained network is then used as a frame-wise feature extractor in order to generate input for an RNN.

The evaluation criteria for the AVEC Emotion Challenge are dependent upon an estimation of the subject's arousal and valence states as a function of time. Various fusion approaches [34, 35] have been applied to this challenge, including state space approach such as Kalman Filtering [12-14] and Particle Filtering [14]. Kalman Filters are the optimal solution to the recursive linear systems estimation problem where process and measurement noise are Gaussian [11], and are utilized in our score fusion approach.

1.1 RECOLA Dataset

The Remote Collaborative and Affective Interactions (RECOLA) database [15] provides the dataset for the AVEC 2016 Emotion Challenge [10]. The corpus contains multimodal signals–audio, video, electro-cardiogram (ECG), and electro-dermal activity (EDA)–recorded synchronously from 27 French-speaking subjects. The subjects have French, Italian or German nationalities to provide some diversity in the expression of emotion. The 27 subjects were broken into three groups of 9 different subjects each: a train (TRAIN) set, a development (DEV) set, and a test (TEST) set.

Ground-truth labeling of the corpus has been performed by six gender balanced French-speaking assistants. Time-continuous ratings of emotional arousal and valence measures are recorded using a 40-msec frame. The corpus provides inter-rater reliability measured by the intra-class correlation coefficient and the Cronbach's α. Ratings are concatenated over all subjects. The root-mean-square error (RMSE), the Pearson Correlation Coefficient (CC), and the Concordant Correlation Coefficient (CCC) values are averaged over all possible pair of raters. In particular, the CCC is chosen as the emotion challenge measure

The rest of this paper is organized as follows. Section 2 provides an overview of the system architecture. Next, we present technical overviews of our audio (Section 3), video (Section 4), physiological (5), and fusion (Section 6) approaches. This includes descriptions of our data processing pipelines, features, and machine learning approaches for training arousal and valence regressors. Section 7 reports our results for an evaluation on the AVEC Emotion Challenge development set and makes comparisons with the AVEC baseline results. Section 8 provides concluding remarks for this work.

2. MITLL-UIUC AVEC ARCHITECTURE

An architectural overview for the channel-level processing of our emotion recognition system for AVEC 2016 is illustrated in Figure 1. Our approach is to integrate multiple machine learning pipelines as well as several different data input processes. Our system takes as input precomputed audio, video, and physiological features from the AVEC 2016 corpus.

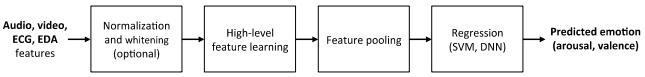

Figure 1. Emotion Recognition System Architecture

Sections 3-5 will discuss the specifics of each of these pipelines, while Section 6 will discuss how to fuse the outputs of each of these systems into a single fused estimate of emotional state.

3. AUDIO PROCESSING

3.1 Overview

For the audio channel, four different sets of audio features were considered for our system, which are discussed later in this section:

1. Baseline AVEC features [10].
2. MFCC features.
3. SDC features
4. Prosody-based audio features.

For each of these audio feature sets, higher-level features are extracted. Our approach is principled in statistical machine learning and discussed in greater detail later in this section. In particular, we employ SVM-based regression. We have implemented high-level feature learning, namely sparse coding, on both the precomputed and extracted multimodal features. This is due to our hypothesis that regression on the learned high-level feature vectors should be more beneficial to emotion recognition than regressing directly on the raw features.

Another important aspect of our system is the early and late fusion. Since we allow multiple feature formats, it is natural to integrate these features before a regression algorithm. This is known as early fusion. Also, since we implement multiple regression algorithms, it makes sense to combine their different regression outputs in a complementary way for the overall improvement in prediction. This late fusion comprises our system's post-processing.

3.2 Audio Feature Extraction

3.2.1 MFCC Feature Extraction

MFCCs of speech frames are computed using a mel-scale filterbank. We extract 20-dimensional cepstral coefficients with a sliding Hamming window that takes in a 20-msec speech frame. The Hamming window is shifted forward with a 10-msec frame rate, resulting in a 50% overlap between the consecutive windowed frames. In addition, we extract 20-dimensional delta cepstral coefficients. The final feature vectors have 40 dimensions formed by stacking the cepstral and delta cepstral coefficients.

3.2.2 SDC Feature Extraction

We perform the shifted delta cepstral feature extraction using a spectral-based technique by Torres et al. [28]. Speech is analyzed with a Hamming window of 20-msec duration at a 10-msec frame rate. The windowed speech waveforms pass through a mel-scale filterbank and RASTA filtering with per-utterance normalization to zero mean and unit variance. The SDC coefficients are calculated using the 7-1-3-7 scheme [19]. Concatenating with static cepstra, the spectral features extracted from speech form a 56-dimensional vector.

3.2.3 Prosody Feature Extraction

Using our audio/speech tools, we extract audio features from the wave files provided in the corpus. The audio preprocessing used in acoustic feature extraction involved estimation of low-level crosstalk in the signal. To isolate regions in which the person of interest is speaking, a simple energy-based method was used as follows. First, the absolute value of the signal is raised to the 1/3 power (an approximation to auditory loudness processing used Todd and Brown [29]) and convolved with a 100 ms Gaussian window. The result is normalized to have a maximum value of 1 and, after informal analysis a threshold of 0.45 was applied to indicate low-energy regions of the audio. Finally, the cross talk regions were estimated as contiguous regions of detected low-energy that were greater than 150ms. Subsequent acoustic analyses of the individual's speech do not consider these crosstalk regions. Rather, for feature time instants that lie within these crosstalk regions, nearest neighbor interpolation was performed.

The acoustic analyses follow a simple, interpretable framework similar to the ideas in [32]. Features are based on vocal effort, variations in intonation and speaking rate. First, vocal effort is captured by loudness and spectral tilt. The loudness is the total loudness output from the perceptual evaluation of audio quality (PEAQ) standard [25]. The spectral tilt is captured with the low order cepstral coefficients (CC0, CC1, and CC2) from a True envelope analysis [26]. The corresponding features are the mean loudness and cepstral coefficients in a 3 second trailing time interval with a 40ms step (to match challenge scoring conditions). Second, variations in intonation are captured by the range and standard deviation of pitch within these 3 second trailing analysis windows. The pitch is extracted using Praat and the top and bottom 5% of the values are removed to mitigate doubling and halving effects. The range (Rf0) and standard deviation (Sf0) of the (log) pitch form the intonation variation features. Finally, in the absence of phonetic alignments, an acoustic measure for speaking rate was estimated by counting the mean number of peak nonstationarities over the 3 second trailing window intervals. Peak nonstationarities are detected from the measure described in [27], smoothed with an 80 ms Gaussian window (to limit any variation within individual phones, e.g. sub 50ms). Together, the loudness, low order cepstral, pitch variation and acoustic speaking rate features represent a set of simple, interpretable measures that inform the emotion prediction.

3.3 High-level audio feature learning

We adopt a semi-supervised approach that uses an unsupervised method, namely sparse coding, followed by a rather simplistic linear regression. The premise of unsupervised learning is to figure out a useful representational mapping by running through unlabeled and unbiased (e.g., uniform mix of various ground-truth labels) examples. To avoid overfitting resulting from inevitably many learned features; we perform max or average pooling before regression.

3.3.1 Sparse coding

Sparse coding aims to learn an efficient data representation using a small number of basis vectors. Given a data input $\mathbf{x} \in \mathrm{R}^N$, sparse coding solves a representation $\mathbf{y} \in \mathrm{R}^K$ (i.e., sparse feature vector of $\mathbf{x}$) while simultaneously updating the dictionary $\mathbf{D} \in \mathrm{R}^{N\mathrm{x}K}$ of K basis vectors in the L1-regularized optimization:

$$\min_{D,y} \|x - Dy\|_2^2 + \lambda \|y\|_1 \quad s.t. \|d_i\|_2 \le 1, \forall i$$

where $\mathbf{d}_i$ is the ith dictionary atom in $\mathbf{D}$, and λ is the penalty that induces a sparse solution $\mathbf{y}$ for a given $\mathbf{x}$. We note that the sparse coding dictionary is an overcomplete matrix, meaning $K > N$. Hence, the solution $\mathbf{y}$ is larger in dimensionality than the input $\mathbf{x}$, but only $S << N$ elements in $\mathbf{y}$ are nonzeros. Sparse coding can alternatively be based on the L0-regularization, although the optimization problem that minimizes the L0 pseudo-norm of a solution in general is known to be intractable. We use the least angle regression (LARS) algorithm for solving the sparse coding problem and Mairal's online dictionary learning method [30].

3.4 Regression methods

We use a linear support vector machine (SVM) to perform the regression task for arousal and valence. Under this regression framework, we optimize the following:

$$\min \tfrac{1}{2}\|w\|^2$$

$$s.t. \quad t - \langle w, x \rangle - b \le \xi \quad and \quad \langle w, x \rangle + b - t \le \xi$$

where $\mathbf{w}$ is the regression weight applied to an input $\mathbf{x}$ for regression target t within a margin parameter ξ. Note the bias unit $\mathbf{b}$ for the regression. Specifically, we consider the L2-regularized L2-loss linear SVM with a unit bias. The SVM complexity parameter has been chosen between 10^{-5} and 1.

We also use support vector regression (SVR). There are two commonly used versions of SVM regression, namely ε-SVR and ν-SVR. The original SVM formulations for regression use the cost parameter C with penalty ε for the points that are incorrectly predicted. An alternative version of the SVM regression applies a slightly different penalty ν. The ν value represents an upper bound on the fraction of training examples that are errors (significantly deviated predictions) and a lower bound for the support vector data points. Nevertheless, the same optimization problem is solved for either case. We have empirically decided to go with ε-SVR.

4. VIDEO PROCESSING

4.1 Regression using video data

4.1.1 Single Frame Regression CNN

We first train a CNN on a single frame to regress the output label. The CNN has 3 convolutional layers consisting of 64, 128, and 256 filters respectively, each of size 5x5. The first two convolutional layers are followed by a 2x2 max pooling while the third layer is followed by quadrant pooling. After the convolutional layers is a fully-connected layer with 300 hidden units and a linear regression layer to estimate the arousal/valence label. All layers save the last one use a rectified linear unit (ReLU) as the nonlinearity function. Our cost function is the mean squared error (MSE). All of our CNNs were trained using stochastic gradient descent with batch size of 128, momentum equal to 0.9, weight decay of 1e-5, and a constant learning of 0.01. All of our CNN models were trained using the anna software library[2].

Prior to passing the video frame to the CNN, we first detect the face in each video frame using face and landmark detector in Dlib-ml [8]. Frames where the face was not detected were dropped. Their scores are later computed by linearly interpolating the scores from adjacent frames. We then use the detected landmarks to normalize the eye and nose locations across faces. We apply mean subtraction and contrast normalization prior to passing each face image through the CNN.

4.1.2 CNN Features as input to an RNN

In order to incorporate temporal information, we have the CNN act as a feature extractor for each video frame and use the resulting feature representation as inputs to an RNN. Specifically, we fix all of the CNN parameters and remove the regression layer. This way, when a frame is passed to the CNN, we extract a 300 dimensional vector from the fully-connected layer. Then, for a given time t, we consider T frames from the past (i.e. [t-T, t]). We pass each frame from time t-T to t to the CNN and extract T vectors in total, each of size 300 dimensions. Each of these vectors is then passed as input (xt) to a node (ht) of the RNN. The hidden state (ht) is computed as the sum of the input via the input weight matrix (Wx) and the previous hidden state via the recurrent matrix (Wh) and a bias (b). The sum is then passed through a nonlinearity (f). Each hidden state in the RNN then regresses the output label (o_t). Once again we use the mean squared error (MSE) as our cost function during optimization.

$$h_t = f(W_x x_t + W_h h_{t-1} + b)$$
$$o_t = W_o h_t$$

Our CNN+RNN model has a single layer RNN with 100 hidden units and a temporal window of size T=100 frames. The model we use for predicting arousal initializes its weights using a Normal distribution, has biases equal to 0, and uses a hyperbolic tangent (tanh) nonlinearity. In contrast, our model for predicting valence initializes its weights using a Uniform distribution, has no bias, and uses a rectified linear unit (ReLU) nonlinearity.

Like our single frame CNN models, our RNN models are trained using stochastic gradient descent with a constant learning rate of 0.01, a batch size of 128 and momentum equal to 0.9. All of the RNNs in our experiments were trained using the Lasagne library.[3]

5. PHYSIOLOGICAL SENSORS

When considering the physiological sensor modalities (ECG, HRHRV, EDA, SCR, SCL), we used the features extracted by the challenge organizers [10]. We elected to use the provided baseline predictions for the ECG, SCR, and SCL features and for the HRHRV and EDA features we trained a Long Short Term Memory network (LSTM) [31], to perform the regression operation.

An LSTM is comprised of a series of cells, each of which has an internal state (c_t) that is updated based on the current input (x_t) and the previous cell state (c_{t-1}). The network then determines how much the previous cell state and the current input contribute to the new cell state using gates. The forget gate (f_t) calculates a value between 0 and 1 using a sigmoid function (σ), which determines the contribution of the previous cell state (c_{t-1}). The input gate (i_t) performs the same operation, but for the current input (x_t). The equations for these operations are shown below:

$$i_t = \sigma(W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f)$$
$$\tilde{c}_t = \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c)$$
$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$$

The model then uses the cell state (c_t) to compute its output representation for time t (h_t). The current cell state's contribution is determined by an output gate (o_t).

$$o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o)$$
$$h_t = o_t * \tanh(c_t)$$

In our experiments, we trained single layer LSTM networks. For the HRHRV features, our networks had 50 hidden units and used a window length of 10 samples. Our arousal model normalized the input data on a per-subject basis and used a constant learning rate of 0.01, while our valence model normalized the input data using all of the subjects in the training set and also had a constant learning rate of 0.001.

For the EDA features, our models, once again, had 50 hidden units. Both models normalized the input data on a per-subject basis and used a constant learning rate of 0.01. Our arousal model had a window length of 10 samples while our valence model had a window length of 50 samples.

All of the models were trained using stochastic gradient descent with momentum. We used a batch size of 128 and momentum value of 0.9. All of the LSTM models were trained using the Lasagne library.[3]

6. MULTIMODAL FUSION

The previous audio, video, and physiological sections discussed distinct and complementary approaches for estimating emotional

[2] https://github.com/ifp-uiuc/anna

[3] https://github.com/Lasagne/Lasagne

state as a function of time. Fusing those emotional measures [34, 35] into a single fused measure is important for improving overall performance and providing robustness in time regions where any given sensor may be faulty or does not provide meaningful information. For example, there are regions where the face is not visible to support feature extraction for the video modality; regions where the person is not speaking to support the audio modality; and instances where there is poor contact for various physiological modalities. Our multimodal fusion approach is used to combine the estimates from these individual channels AND exploit the time-series nature of the data. Our approach leverages a Kalman Filter-based approach [11] for estimating the emotional state (x) as a function of time from the information (z) from the respective channels using the standard state space framework. The state transition equation models the time-varying nature of the emotional states, where *A* is the transition matrix and *w(k)* is the zero-mean process noise perturbing the system:

$$x(k+1) = Ax(k) + w(k)$$

The measurement equation relates how the measures (z) from the individual measurement channels relate to the underlying emotional state (*x*):

$$z(k) = Cx(k) + \beta + v(k) = \begin{bmatrix} z_{audio}(k) \\ z_{video\_appearance}(k) \\ z_{video\_geometric}(k) \\ z_{physiological}(k) \\ \vdots \end{bmatrix}$$

The measurement matrix (*C*) relates the underlying emotional states to the measurements and *v(k)* is the zero-mean measurement noise term. In practice, we found that the measurement noise was often non-zero, so a bias term (β) has been added to the model, which has proven useful to the AVEC problem.

Held out data is used to perform the system identification problem of determining the system matrices and noise terms. Held out data from the TRAIN and/or DEV sets are used to model x from the gold standard (annotated truth for the emotions) and z from the corresponding measurements from the individual channels:

$$X_{1,N} = \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix}$$

$$Z_{1,N} = \begin{bmatrix} z_1 & \cdots & z_N \end{bmatrix}$$

In some cases the different z's may correspond to different sensor channels, though may also be different models for the same sensor channel (e.g. audio MFCC and audio SDC). For example, x_m would correspond to a scalar value representing the emotional state (e.g. arousal, valence) for sample m, while z_m would correspond to a vector of emotional state measurements corresponding to each of the applied sensor channels/models. This enables us to model the state transition matrix (*A*) and the variance of the process noise (*Q*):

$$A = \left(X_{2,N} X_{1,N-1}{}^T\right)\left(X_{1,N-1} X_{1,N-1}{}^T\right)^{-1}$$

$$Q = \mathrm{cov}(w,w) = \mathrm{cov}\left(X_{2,N} - AX_{1,N-1}\right)$$

If we make the following substitutions:

$$\overline{X}_{1,N} = \begin{bmatrix} X_{1,N} \\ 1_{1xN} \end{bmatrix}$$

$$\overline{C} = \begin{bmatrix} C & \beta \end{bmatrix}$$

We can rewrite the measurement equation as follows:

$$Z_{1,N} = \overline{C}\,\overline{X}_{1,N} + v(k)$$

This enables a convenient form factor for deriving the measurement matrix (*C*), the bias term (β), and the variance of the measurement noise (*R*):

$$\overline{C} = \left(Z_{1,N} \overline{X}_{1,N}{}^T\right)\left(\overline{X}_{1,N} \overline{X}_{1,N}{}^T\right)^{-1}$$

$$R = \mathrm{cov}(v,v) = \mathrm{cov}\left(Z_{1,N} - CX_{1,N} - \beta\right)$$

This model provides a useful approach for fusing sensor channel measurements per time step, and also models the time-varying nature of the model to further improve system performance. Backward smoothing [11] was also used to further improve system performance by leveraging *future* measurements to improve current estimates.

7. RESULTS

7.1 Dataset and Evaluation

The experiments in this section use the RECOLA dataset [15] for training and evaluation. The experiments in this section use the evaluation protocol defined in the AVEC 2016 Emotion Challenge [10]. Our models were trained on the provided TRAIN set and were evaluated on the DEV set. Note that baseline system results are available for this evaluation paradigm [10], as listed in Table 1.

Table 1. Baseline CCC results [10] for AVEC Emotion Challenge on the DEV Set

Modality	Arousal	Valence
Audio	0.796	0.455
Video (appearance)	0.483	0.474
Video (geometric)	0.379	0.612
ECG (electrocardiogram)	0.288	0.153
HRHRV (heart rate & heart rate variability)	0.382	0.293
EDA (electrodermal activity)	0.077	0.104
SCL (skin conductance level)	0.101	0.124
SCR (skin conductance resistance)	0.071	0.110
Multimodal	0.820	0.702

7.2 Audio Results

We report the emotion prediction performance by the audio features only. We have achieved particularly strong performance from the MFCC and SDC feature sets, both of which are followed by sparse coding. These results exceed the arousal score for the baseline system. For valence, we also achieve the best performance for the MFCC and SDC feature sets. We have

optimized sparse coding for SDC using 256 to 512 dictionary atoms with the regularization parameter $\lambda = 0.2$. A linear kernel was used for the SVM that performed the back end regression.

Table 2. Performance on AVEC 2016 DEV Set Using only Audio Features

		RMSE	PCC	CCC
Arousal	Baseline Features	0.138	0.771	0.751
	MFCC Features	0.107	0.846	0.846
	SDC Features	0.123	0.807	0.800
	Prosody Features	0.186	0.718	0.608
Valence	Baseline Features	0.135	0.441	0.433
	MFCC Features	0.132	0.456	0.450
	SDC Features	0.133	0.445	0.443

7.3 Video Results

In Table 3, we show how well our CNN+RNN architecture performs at predicting the arousal and valence scores of subjects in the DEV set. We see that the CNN+RNN does a much better job at predicting valence than arousal. This is not surprising as many previous works have shown this to be the case. We also see that our learned CNN+RNN feature representation outperforms the baseline trained on handcrafted video appearance features (CCC=0.511 vs. CCC=0.474).

Table 3. CNN+RNN Performance for video appearance on the AVEC 2016 DEV Set

	RMSE	PCC	CCC
CNN+RNN (arousal)	0.201	0.415	0.346
CNN+RNN (valence)	0.107	0.549	0.511

7.4 Physiological Results

We report the performance on the DEV set for our LSTM models trained on the HRHRV and EDA features in Table 4. When generating our predictions, we employed the same post-processing pipeline used by the challenge organizers [10] which is described in [33]. It consists of (i) smoothing with a median filer (ii) centering (iii) scaling and, in the case of the EDA features, (iv) time-shifting the predictions. Each post-processing step was kept and applied to the TEST set if it improved the CCC score on the DEV set. We see that by using an LSTM we achieve comparable performance with baseline when estimating arousal and improve performance considerably when estimating valence. The reason for the marked improvement in valence may be due to the fact that predicting valence requires more temporal information (longer window lengths), thus, having a model that explicitly models the temporal dynamics of the features (LSTM) is preferable to a model that considers the time window all at once.

Table 4. LSTM Performance for physiological sensors on the AVEC 2016 DEV Set.

		RMSE	PCC	CCC
Arousal	HRHRV	0.218	0.407	0.357
	EDA	0.250	0.089	0.082
Valence	HRHRV	0.117	0.412	0.364
	EDA	0.124	0.267	0.177

7.5 Multimodal Results

Our multimodal system fuses the emotional states derived from the individual audio, video, and physiological sensors discussed in the previous subsections using the Kalman Filter framework discussed in Section 6. Models for the transition matrix (A), measurement matrix (C), measurement bias (β), process noise (Q), and measurement noise (R) are estimated from the TRAIN and DEV set subjects. (For DEV set evaluation we have 9 partitions of the DEV subjects where we hold out the subject under evaluation and use the remaining DEV subjects and all of the TRAIN subjects.) Backward smoothing was found to improve performance, as did the bias compensation for the individual channels. The channels fused for arousal and valence for the multimodal system include the feature channels discussed in Sections 3-5, as well as the AVEC baseline features [10] for video appearance, video geometric, and ECG. We also included a sparse coding backend to the baseline video geometric system, as we did for the audio channels as discussed in Section 3.

The arousal and valence results for our multimodal systems are contained in Table 5 for both DEV set and TEST set data. The DEV set results are self-reported, while the TEST set results are official results from the AVEC Evaluation. For comparison, baseline system performance results [10] are also included in Table 5.

The multimodal results exhibit meaningful improvements over the unimodal results, particularly for valence. They also demonstrate significant performance results over the baseline cases for both arousal and valence on both the DEV set and TEST set partitions.

Table 5. Multimodal results on the DEV and TEST sets, including MITLL-UIUC and Baseline scores

		RMSE	PCC	CCC (Baseline)
Arousal	DEV SET	0.103	0.862	0.862 (0.820)
	TEST SET	0.115	0.774	0.770 (0.702)
Valence	DEV SET	0.089	0.751	0.750 (0.682)
	TEST SET	0.100	0.689	0.687 (0.638)

8. CONCLUDING REMARKS

This paper provided an overview of our AVEC 2016 Emotion Challenge technical approaches and corresponding results that exceeded the CCC baseline results on the TEST set. The MFCC and SDC audio approaches with sparse coding backends provided significant performance improvements for arousal on the DEV set over the baseline scores. Likewise, the deeply learned models for video appearance, HRHRV, and EDA provided significant performance improvements for valence on the DEV set over the baseline scores. The fusion approach enabled the multi-sensor fusion of emotional state while leveraging the time-varying nature of the emotional states.

Near term work includes further refinement to the individual sensor channel approaches introduced in this paper. It will also include an improved noise model to account for the non-stationary nature of the noise in the various sensor channels. This would include the exploitation of speech activity detection (SAD) and adjusting the video noise model where the face is unobservable.

9. REFERENCES

[1] L. Chao, J. Tao, M. Yang, Y. Li, & Z. Wen, 2015. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge,* 65-72. ACM.

[2] S. Chen, & Q. Jin, 2015. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 49-56. ACM.

[3] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, & C. Pal, 2015. Recurrent neural networks for emotion recognition in video. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 467-474. ACM.s

[4] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, & H. Sahli, 2015. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*,73-80. ACM.

[5] S.E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R.C. Ferrari, and M. Mirza, 2013, December. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, 543-550. ACM.

[6] P. Khorrami, T. L. Paine, K. Brady, C. Dagli, & T. S. Huang, 2016. How Deep Neural Networks Can Improve Emotion Recognition on Video Data. *arXiv preprint arXiv:1602.07377*.

[7] P. Khorrami, T. Paine, & T. Huang, 2015. Do deep neural networks learn facial action units when doing expression recognition?. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 19-27.

[8] D. E. King, 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 1755-1758.

[9] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. Pattern Recognition Letters, 66, 22-30.

[10] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge}, *arXiv:1605.01600*, 2016.

[11] Y. Bar-Shalom, X. Rong Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory, Algorithms, and Software*, John Wiley & Sons, Inc., New York, 2001.

[12] M. Glodek, et al, *Kalman Filter Based Classifier Fusion for Affective State Recognition*, Multiple Classifier Systems, Vol. 2772 of the series Lecture Notes in Computer Science, Springer-Verlag, Berlin,, 2013.

[13] M. Kachele, et al, *Fusion of Audio-Visual Features Using Hierarchical Classifier Systems for the Recognition of Affective States and the State of Depression*, Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM), 2014.

[14] K. Markov, et al, *Dynamic Speech Emotion Recognition with State-Space Models*, 23rd European signal Processing Conference (EUSIPCO), 2015.

[15] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. *Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions*. In Proc. of EmoSPACE, FG, Shanghai, China, 2013.

[16] M. Grimm, K. Kroschel, and S. Narayanan. Support vector regression for automatic recognition of spontaneous emotions in speech. In *Proc. of IEEE International Conference on Acoustics, Speech, Speech and Signal Processing*, 2007.

[17] M. Wöllmer et al., Abandoning emotion classes-towards continuous emotion recognition with modeling of long-range dependencies. In *Proc. of Interspeech*, 2008.

[18] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data," in *Proc. of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, ACM MM, Brisbane, Australia, October 2015.

[19] Campbell, W. M., Singer, E., Torres-Carrasquillo, P. A., Reynolds, D. A., Language Recognition with Support Vector Machines, *In Proc. Odyssey: The Speaker and Language Recognition Workshop* in Toledo, Spain, ISCA, pp. 41–44, 31 May–3 June 2004.

[20] Z. Huang et al. An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction. In *Proc. of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, ACM MM, Brisbane, Australia, October 2015.

[21] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[22] B. A. Olshausen and D. J. Field. Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1? *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[23] A. Coates, H. Lee, A. Ng. An analysis of single-layer networks in unsupervised feature learning. In Proc. of *AISTATS*, 2011

[24] R. Rigamonti, M. A. Brown and V. Lepetit. Are sparse representations really relevant for image classification? In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2011

[25] ITU Standard rec-bs.1387-1-2001, 2001.

[26] A. Roebel and X. Rodet, “Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation,” in Digital Audio Effects (DAFx), 2005, pp. 30–35.

[27] D. Kapilow, Y. Stylianou, and J. Schroeter, Detection of nonstationarity in speech signals and its application to time-scaling. *Eurospeech*. 1999. pp. 2307–2310.

[28] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller. Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features. In *Proc. of INTERSPEECH*, 2002.

[29] N. Todd and G. Brown. Visualization of Rhythm, Time and Metre. *Artificial Intelligence Review*, vol. 10, pp. 253–273, 1996

[30] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online Dictionary Learning for Sparse Coding. In *Proc. of ICML*, 2009

[31] A. Graves. "Generating sequences with recurrent neural networks." *arXiv preprint arXiv:1308.0850* (2013).

[32] D. Bone, C. Lee, and S. Narayanan, “Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features,” *Affective Computing, IEEE Transactions on*, vol. 5, no. 2, pp. 201–213, 2014.

[33] G. Trigeorgis F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, and S. Zafeiriou. "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network." In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200-5204. IEEE, 2016.

[34] Glodek M, Honold F, Geier T, Krell G, Nothdurft F, Reuter S, Schüssel F, Hörnle T, Dietmayer K, Minker W, Biundo S, Weber M, Palm G, Schwenker F (2015) Fusion paradigms in cognitive technical systems for human-computer interaction. Neurocomputing 161:17–37.

[35] M. Kachele, M. Schels, P. Thiam and F. Schwenker: "*Fusion Mappings for Multimodal Affect Recognition*", in proc. of IEEE Symposium Series on Computational Intelligence, pp. 307-313, Cape Town, South Africa, December 7-10 2015.