Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks

Lang He
NPU-VUB Joint AVSP Lab
School of Computer Science
Northwestern Polytechnical University (NPU)
127 Youyi Xilu, Xi'an 710072, China
langhe@mail.nwpu.edu.cn

Dongmei Jiang
NPU-VUB Joint AVSP Lab
School of Computer Science
Northwestern Polytechnical University(NPU)
127 Youyi Xilu, Xi'an 710072, China
jiangdm@nwpu.edu.cn

Le Yang
NPU-VUB Joint AVSP Lab
School of Computer Science
Northwestern Polytechnical University(NPU)
127 Youyi Xilu, Xi'an 710072, China
yangle.cst@gmail.com

Ercheng Pei
NPU-VUB Joint AVSP Lab
School of Computer Science
Northwestern Polytechnical University(NPU)
127 Youyi Xilu, Xi'an 710072, China
peierch@mail.nwpu.edu.cn

Peng Wu
NPU-VUB Joint AVSP Lab
Dept. Electronics & Informatics (ETRO)
Vrije Universiteit Brussel(VUB)
Pleinlaan 2, 1050 Brussels, Belgium
pwu@etro.vub.ac.be

Hichem Sahli
NPU-VUB Joint AVSP Lab
Dept. ETRO, VUB
Pleinlaan 2, 1050 Brussels
Interuniversity Microelectronics Centre
Kepeldreef 75, 3001 Heverlee, Belgium
hsahli@vub.ac.be

ABSTRACT

This paper presents our system design for the Audio-Visual Emotion Challenge (AV^+EC 2015). Besides the baseline features, we extract from audio the functionals on low-level descriptors (LLDs) obtained via the YAAFE toolbox, and from video the Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) features. From the physiological signals, we extract 52 electro-cardiogram (ECG) features and 22 electro-dermal activity (EDA) features from various analysis domains. The extracted features along with the AV^+EC 2015 baseline features of audio, ECG or EDA are concatenated for a further feature selection step, in which the concordance correlation coefficient (CCC), instead of the usual Pearson correlation coefficient (CC), has been used as objective function. In addition, offsets between the features and the arousal/valence labels are considered in both feature selection and modeling of the affective dimensions. For the fusion of multimodal features, we propose a Deep Bidirectional Long Short-Term Memory Recurrent Neural Network (DBLSTM-RNN) based multimodal affect prediction framework, in which the initial predictions from the single modalities via the DBLSTM-RNNs are firstly smoothed with Gaussian smoothing, then input into a second layer of DBLSTM-RNN for the final prediction of affective state. Experimental results show that our proposed features and the DBLSTM-RNN based fusion framework obtain very promising results. On the development set, the obtained CCC is up to 0.824 for arousal and 0.688 for valence, and on the test set, the CCC is 0.747 for arousal and 0.609 for valence.

AVEC'15, October 26, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2808196.2811641.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*signal processing, computer vision*

Keywords

DBLSTM-RNN; multimodal fusion; offset; physiological feature; audio and video features

1. INTRODUCTION

In recent years, recognition of non-acted spontaneous emotions in the continuous dimensional space has attracted researchers' interest. By providing the audio visual multimodal emotion databases along with the continuous labels, the Audio-Visual Emotion Challenge (AVEC2011 through AVEC2014 [26], [27], [32], [31]) helped speeding up the development of new frameworks for continuous affect recognition, which are summarized here after.

From the audio signals, AVEC organizers provided baseline features as functionals on the low-level descriptors (LLDs), such as loudness, zero crossing rate, spectral flux, Mel-frequency cepstral coefficients (MFCCs), and voicing related features such as jitter, shimmer, logarithmic Harmonics-to-Noise Ratio (logHNR) *etc.* The adopted functionals are statistical functionals, regression functionals, and local minima/maxima related functionals *etc.* The above features have been proved efficient in predicting the affective dimensions, especially

arousal. Among other features, Meng *et al.* in [17] performed Motion History Histograms (MHH) on the LLDs to extract change information of the vocal expressions. In [18], Mitra *et al.* explored a wide array of acoustic features that capture speech articulation, acoustic-phonetic information, spectral representation, speech modulation, vocal effort, rhythmicity, speech prosody, vowel stress, *etc.* Experimental results showed that these features obtain very promising performance in estimating the depression scales.

For the video features, AVEC provided the uniform Local Binary Patterns (LBP) features, Local Phase Quantization (LPQ) features, and Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) features in different challenge years. Besides these features, in [20] and [2], the authors also extracted face shape features, global appearance features and local appearance features. In [17], Meng *et al.* first computed MHH from the continuous image sequences, then highlighted its temporal details by Edge Orientation Histogram (EOH) and LBP. While in [13], they first extracted the EOH, LBP and LPQ features, then used MHH to capture the dynamic movement of the features. In [10], besides the LGBP-TOP features, the optical-flow-based motion vectors as well as facial landmark features were extracted, too.

Physiological features, such as electro-cardiogram (ECG) and electro-dermal activity (EDA) features, have also been extracted from time domain and frequency domain [30] [14] for emotion recognition. Heart rate and heart rate variability are the most often reported emotion indicators, followed by skin conductance level [15]. Besides these features, in [9], Guo computed statistics from the time-frequency representation of the signal provided by wavelet transform. In [12] and [1], the authors extracted features based on empirical mode decomposition, which relies on a fully data-driven mechanism that is especially well suited for non-linear signals. Non-linear features, such as Hurst exponent [3], approximate entropy [28], sample entropy [14], and dominant Lyapunov exponent [29] have also been used as they correlate with emotion variation.

With respect to the continuous affect recognition, various regression models have been proposed, the reader is referred to Gunes *et al.* [11] for a survey. The most commonly used models are static regression models, such as kernel based Support Vector Regression (SVR) [20] and Relevance Vector Regression (RVR), or linear regression models [33]. These regression models are effective in continuous affect recognition but are insufficient in capturing the temporal information of the affective dimensions. Wöllmer *et al.* in [39] used (unidirectional) Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) to perform regression analysis on valence and arousal. LSTM-RNN, being able to learn the long-range temporal dependencies, demonstrated the benefits of including temporal modeling. Later, Wöllmer and Nicolaou *et al.* further improved the LSTM architectures to bidirectional LSTM (BLSTM) by including the *forget gates* and bidirectional processing [40] [19] [35]. BLSTM-RNN can learn when to store or relate to context information over long periods of time, while the application of non-linear functions enables learning non-linear dependencies. Currently the LSTM(BLSTM)-RNN model is still one of the best regression models obtaining remarkable performance in affective computing.

Significant improvements have been obtained on multimodal affect recognition. Nicolle *et al.* [20] fused the initial estimation results from the Nadaraya-Watson kernel regressors via local linear regression (LLR). In [10], Gupta *et al.* proposed a multi-layered system, in which the initial predictions, after being smoothed by a moving average filter, are fused via linear regression (LR), and the fused results are further processed using a temporal regression. In [2], a Deep Belief Network -Linear Regression (DBN-LR) fusion framework was proposed. The extracted features are input into a Restricted Boltzmann Machine (RBM) with one hidden layer, then a temporal pooling function is added, and the pooled features are fed into the linear regression layer instead of a Multiple Layer Perceptron (MLP) for the regression of the affective dimension. Predictions of affective dimensions from all the signals (in the same window size) are input to the LR model for final decision. In [19], Nicolaou *et al.* proposed the BLSTM-RNN based output-associative fusion framework, in which the valence and arousal values from the original input features via a first layer of BLSTM-RNNs are jointly input into a second layer BLSTM-RNN to get the final predictions of valence (arousal). This framework has the advantage of modeling the correlation between the dimensions.

Deep neutral networks (DNNs), having the strong ability of capturing the underlying nonlinear relationship among data with multiple layers, have been successfully used in many fields. However, they can only provide limited temporal modeling by operating on a fixed-size sliding window of feature frames. By stacking multiple BLSTM-RNN layers, Deep BLSTM-RNN (DBLSTM-RNN), which takes both the advantage of DNN with deep structures, and the advantage of BLSTM of modeling the long term temporal history, has been designed and obtained great success in speech recognition [25, 6] and speech synthesis [5].

In this paper, we target the multi-modal affect recognition task of AV^+EC 2015 [23]. The contributions are three folds:

1) Referring to the BLSTM-RNN based output associative fusion framework of [19], we propose a DBLSTM-RNN based hybrid multi-modal affect recognition framework, in which the initial predictions from the single modalities via the DBLSTM-RNNs are firstly smoothed with Gaussian smoothing, then input into a second layer of DBLSTM-RNN for the final prediction of affective state.

2) Besides the baseline features, we estimate extra features. From the audio streams, we perform functionals on the low-level descriptors (LLDs) obtained via the YAAFE toolbox [16]. From video, LPQ-TOP features are extracted. Finally 52 ECG features as well as 22 EDA features are extracted from the physiological signals from various analysis domains.

3) A feature selection scheme, on the concatenated features, has been devised. Moreover, before feature selection the concordance correlation coefficient (CCC) has been used to estimate the time offset between the features and the arousal/valence labels. The estimated offset is used as a delay parameter of the input features to the DBLSTM-RNN models.

Experimental results show that our proposed DBLSTM-RNN based fusion framework obtain very promising results.

The remainder of this paper is organized as follows. The extracted audio, video and physiological features, as well as the offset issue and feature selection issue, are addressed in

section 2. The proposed multi-modal affective dimension prediction framework using DBLSTM-RNN is described in section 3. Section 4 analyzes the experimental results and finally conclusions are drawn in section 5.

2. MULTI-MODAL EMOTION FEATURES

The AV^+EC 2015 challenge [23] adopts the RECOLA corpus [24], which was recorded to study socio-affective behaviours from multimodal data in the context of remote collaborative work. Audio, video, ECG and EDA signals were synchronously recorded from 27 French-speaking subjects. The dataset is equally divided in three partitions: train, development and test, each having 9 recordings of 5 minutes together with the arousal and valence ratings of every 40 ms.

2.1 Audio Features

The 102 baseline audio features of AV^+EC 2015 adopt the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) containing functionals on the 42 low-level descriptors (LLDs) [23] which are extracted with the openSMILE toolkit [4]. These LLDs cover the spectral, cepstral, prosodic and voice quality information.

Besides the baseline audio features, we use the YAAFE toolbox [16] to extract other 158 LLDs, as shown in Table 1 where the numbers between brackets are dimensions of the extracted feature vectors, and "OBSI" is the abbreviation of octave band signal intensity. As the Amplitude Modulation and Envelope Shape Statistics features need low frequency information, they are extracted with the frame length of 1 s and frame shift of 10 ms. All the other LLDs are extracted with the frame length of 23 ms and also frame shift of 10 ms. 29 functionals, as shown in Table 2, are computed by openSMILE on these LLDs within the overlapping short segments (3 s) with segment shift of 40 ms, resulting in 4582 dimensional feature vectors.

Finally, we concatenate the 102 baseline features with the 4582 YAAFE based features, resulting in 4684 dimensional feature vectors with the frame rate of 25 frames/s, denoted hereafter as Audio-CON.

Table 1: Low level descriptors from YAAFE (158)

Energy and spectral(119)	
Amplitude Modulation(8)	Energy(1)
Envelope Shape Statistic(4)	Loudness(24)
LPC(10)	MFCC(13)
LPC-d1(10)	MFCC-d1(13)
LPC-d2(10)	MFCC-d2(13)
Perceptual Sharpness(1)	Perceptual Spread(1)
Spectral Decrease(1)	Spectral Flatness(1)
Spectral Flux(1)	Spectral Shape Statistic(4)
Spectral Flux-d1(1)	Spectral Slope(1)
Spectral Flux-d2(1)	Spectral Variation(1)
Voicing related (39)	
OBSI(10)	OBSI-d1(10)
OBSI-d2(10)	OBSI ratio(9)

Table 2: 29 functionals from openSMILE

Statistical functionals (25)
max, min, arithmetic mean, norm, variance, stddev, skewness, kurtosis, numPeaks, meanPeakDist, peakMean, peakMeanMeanDist, quartiles, samplepos, maxNumSeg, rangeRelThreshold, numSegments, meanSegLen, maxSegLen, minSegLen, upleveltime25, upleveltime50, upleveltime75, risetime, falltime
Regression functionals (4)
linregc1, linregc2, linregerrA, linregerrQ

2.2 Video Features

2.2.1 Baseline Video Features

In AV^+EC 2015, the LGBP-TOP features are used as appearance features. Principal Component Analysis (PCA) is performed on the LGBP-TOP features, resulting in 84 dimensional feature vectors with the frame rate of 25 frames/s. The 316 baseline geometric features are calculated on the facial landmarks. For those frames of zero vectors, we use the linear interpolations of the adjacent non-zero frames to replace the zero vectors.

2.2.2 LPQ-TOP Features

Besides the baseline appearance and geometric features, we also extract Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) features. Facial areas are firstly detected according to the face bounding boxes provided by AV^+EC 2015, then the face video is split into spatio-temporal video volumes containing 75 frames (3 seconds window). As extraction of the LPQ-TOP features is time consuming, the window shift is set as 1 second. We calculate the phase information locally for every image position in three directions, and the phase of the four low-frequency coefficients are de-correlated and uniformly quantized in a 24-dimensional space. The resulting binary patterns are histogrammed for the three orthogonal slices separately, denoted as XY-LPQ, XT-LPQ, and YT-LPQ, respectively, and concatenated into a single feature histogram. Just as with the LGBP-TOP features, the X-Y plane provides the spatial domain information while the X-T and Y-T planes provide temporal information. Therefore, by using this dynamic texture descriptor, both appearance and motion in three directions are considered. Finally, 768 (256*3) LPQ-TOP features are extracted per space-time volume.

As the LPQ-TOP feature vectors are extracted every 1 second, to be consistent with the frame rate (25 frames/s) of arousal or valence labels, we adopt cubic spline interpolation [21] on the LPQ-TOP feature vectors.

2.3 Physiological Features

The AV^+EC 2015 challenge provided 54 baseline features from the electro-cardiogram (ECG) signals, and 60 baseline features from the electro-dermal activity (EDA) signals, respectively, with overlapping windows (40 ms shift) of 4s length. The readers are referred to [23] for more details.

In our work, from the two-channel physiological signals, we extract extra 52 ECG features and 22 EDA features, with overlapping windows (40 ms shift) of 4 s length, from various analysis domains including time, frequency, time-frequency, and non-linear domains. In summary, we first compute from both channels, the following general features: a) conventional statistics of the raw signal – mean, standard deviation, maximum, minimum, root mean square, skewness, and kurtosis [30]; b) non-linear features – the approx-

imate entropy, sample entropy, and Hurst exponent [30]; c) statistics of wavelet coefficients – maximum, minimum, standard deviation values of all scales of wavelet coefficient [9]; d) empirical mode decomposition (EMD) based features – the root mean square of the amplitude of the analytic version of all Intrinsic Mode Functions (IMFs), the maximum amplitude of the analytic version of all IMFs, the mean instantaneous frequency of all IMFs, and the weighted mean instantaneous frequency of all IMFs [36]. It is worth noting that we did not extract the wavelet based and EMD based features from EDA signal, as a window of EDA is usually too simple to be decomposed meaningfully.

Besides these features, specific features have been also estimated.

ECG: heart rate, the standard deviation of Normal-to-Normal (NN) intervals, the root mean square value of the differences of successive NN intervals, and the proportion of interval difference of successive NN intervals greater than 50 ms, are obtained as time-domain features [14]. For frequency domain features, we extract the average power of the low frequency band (0.04-0.15 Hz) and the high frequency band (0.15 - 0.4 Hz), as well as the ratio of power within the low frequency band to that within the high frequency band [14].

Skin Conductance (SC): The EDA complex includes both background tonic (skin conductance level: SCL) and rapid phasic components (skin conductance responses: SCRs). Conventional statistics obtained in the SCL analysis have been found to be correlated with emotion. Apart from the above general statistics features, extra statistics are calculated, including: the mean of the first derivation, mean of the second derivation, mean of negative slope. Moreover, the average power in the frequency bands (0-0.1 Hz, 0.1-0.2 Hz, 0.2-0.3 Hz, 0.3-0.4 Hz) are computed [5]. For SCR analysis, we calculate the rate of SCR occurrences from the very low frequency band (0-0.1 Hz) and the low frequency band (0-0.2 Hz), and the ratio between both rates [30].

Finally, we concatenate the extracted features with the baselines features, from ECG and from EDA, respectively, resulting in totally 106 ECG features (denoted hereafter as ECG-CON) and 82 EDA features (denoted hereafter as EDA-CON).

Notice that when we extracted the above ECG and EDA features, we found that 1 video clip in the training set, and 3 video clips out of the whole 9 in the test set were not well recorded. The ECG and EDA signals contain artifacts probably caused by the motion of the subjects. In this case, significant but non-emotion-related changes are present, which may lead to a decreasing performance of our approaches. Therefore, for these video clips, we did not extract the above physiological features.

2.4 Offsets Between Features and Labels

When the annotators labeled the arousal or valence, delays (offsets) might exist between the labels and the video. For each feature type, we use the feature vectors in the training set along with their labels to find the best offset, by calculating the mean Concordance Correlation Coefficient (CCC) of the feature elements. For the dimension n, the CCC between the feature element sequence $Y_n = \{y_{n1}, ..., y_{nT}\}$ and the label sequence $X = \{x_1, ..., x_T\}$ is defined as where ρ is the Pearson correlation coefficient (CC), μ and σ are mean and variance, respectively. The mean

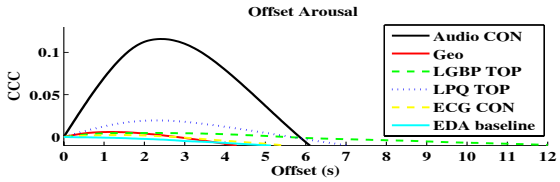

Figure 1: CCC vs. offset for arousal

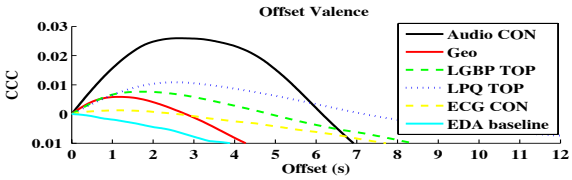

Figure 2: CCC vs. offset for valence

CCC is calculated as the average on all dimensions of the feature vector.

$$CCC = \frac{2\rho\sigma_{Y_n}\sigma_X}{\sigma_{Y_n}^2 + \sigma_X^2 + (\mu_{Y_n} - \mu_X)^2} \quad (1)$$

Figure 1 shows the curves of the mean CCCs (subtracted by the values at offset 0) versus offsets for arousal, those for valence are shown in Figure 2. One can notice that for some features, especially Audio-CON, when the offsets are appropriately considered, the mean CCC increase greatly. Therefore, in feature selection and modeling of the affective dimensions, we synchronize the Audio-CON, LPQ-TOP, geometric features (Geo), and LGBP-TOP with the arousal or valence labels using the offsets listed in Table 3, which correspond to the peak values in Figure 1 and Figure 2.

2.5 Feature Selection

We adopt the correlation based feature selection (CFS) [22] on the above extracted features, with sequential floating forward selection (SFFS) algorithm [34] as the searching strategy. In the CFS algorithm, we use CCC as the objective function. To get more reliable results, features from both the training set and the development set are adopted for feature selection. Notice that feature selection has not been performed on the LGBP-TOP features (their dimensionality has been reduced by PCA), the ECG and EDA features (they have low dimensions). The final dimensions of the selected features are listed in Table 3, where 'FS' means after feature selection, and the number of selected features from our newly proposed features is indicated between brackets. It has to be mentioned that for audio almost all the selected features are the proposed ones (features based on the LLDs from YAAFE), and for EDA features, only 1 of the proposed features has been selected.

3. DBLSTM-RNN BASED AFFECTIVE DIMENSION PREDICTION

3.1 DBLSTM-RNN Model

The main theory of DBLSTM-RNN described here is based on [8], [7], and [38]. Given an input feature sequence $x =$

Table 3: Offsets and dimensions of the feature vectors after feature selection

Modality	Offset		Feature Dimension	
	Arousal	Valence	Arousal	Valence
Audio-CON-FS	59	78	15(15)	27(24)
LPQ-TOP-FS	58	64	36	23
Geo-FS	24	43	10	8
LGBP-TOP	24	43	84	84
ECG-CON-FS	–	–	14(10)	15(5)
EDA-CON-FS	–	–	19(1)	12(1)
ECG baseline	–	–	54	54
EDA baseline	–	–	60	60

$(x_1, ..., x_T)$, a standard long short-term memory recurrent neural network (LSTM-RNN) computes the hidden vector sequence $h = (h_1, ..., h_T)$ and output vector sequence $y = (y_1, ..., y_T)$ by iterating the following equations from $t = 1$ to T:

$$(h_t, c_t) = H(x_t, h_{t-1}, c_{t-1}) \quad (2)$$

$$y_t = W_{hy} h_t + b_y \quad (3)$$

where the H term is the LSTM-RNN layer function, c is the *cell* activation vector with the same size as the hidden vector h. The W terms denote weight matrices (e.g. W_{hy} is the hidden-output weight matrix), and the b terms denote the bias vectors (e.g. b_y is the output bias vector).

Deep LSTM-RNN has the characteristic of DNN. It can be created by stacking multiple LSTM-RNN hidden layers on top of each other, with the output sequence of one layer forming the input sequence for the next. Assuming the deep LSTM-RNN has N hidden layers in the stack, then the hidden vector sequence h^n are iteratively computed from $n = 1$ to N:

$$(h_t^n, c_t^n) = H^n(h_t^{n-1}, h_{t-1}^n, c_{t-1}^n) \quad (4)$$

with $h_t^0 = x_t$. The network output y_t are then computed as

$$y_t = W_{h^N y} h_t^N + b_y \quad (5)$$

DBLSTM-RNN is a bidirectional extension of deep LSTM-RNN. The iteration equations are as follows.

$$(\overrightarrow{h}_t^n, \overrightarrow{c}_t^n) = \overrightarrow{H}^n(h_t^{n-1}, \overrightarrow{h}_{t-1}^n, \overrightarrow{c}_{t-1}^n) \quad (6)$$

$$(\overleftarrow{h}_t^n, \overleftarrow{c}_t^n) = \overleftarrow{H}^n(h_t^{n-1}, \overleftarrow{h}_{t-1}^n, \overleftarrow{c}_{t-1}^n) \quad (7)$$

where $h_t^{n-1} = [\overrightarrow{h}_t^{n-1}; \overleftarrow{h}_t^{n-1}]$, and $\overrightarrow{h}_t^0 = x_t$, $\overleftarrow{h}_t^0 = x_t$. The network output y_t are

$$y_t = W_{\overrightarrow{h}^N y} \overrightarrow{h}_t^N + W_{\overleftarrow{h}^N y} \overleftarrow{h}_t^N + b_y \quad (8)$$

The parameters of W and b can be learned by back-propagation through time from the training data, with the sum of the squared deviations between the outputs $\{\widehat{y_t}\}$ and the ground truth labels $\{y_t\}$ $(t = 1, ..., T)$ as the error function. Details can be found in [8] and [7].

3.2 Multi-Modal Affect Prediction

The DBLSTM-RNN based multimodal affective dimension prediction framework is as shown in Figure 3. In our experiments two of such architecture are used for valence and

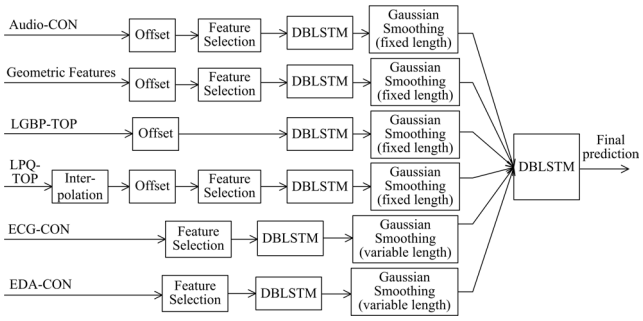

Figure 3: The multi-modal affective dimension prediction framework.

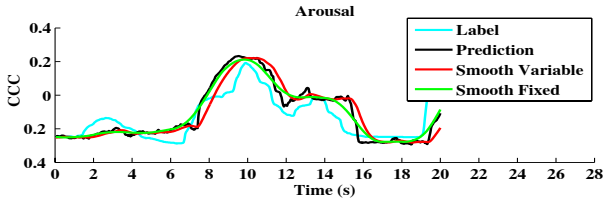

Figure 4: The smoothed arousal predictions.

arousal, respectively. The LPQ-TOP features are firstly interpolated to match the frame rate of the arousal/valence labels, then the Audio-CON, LGBP-TOP, LPQ-TOP and geometric features are delayed with their corresponding offsets as listed in Table 3 (the missing features at the beginning are duplicated from the delayed ones). While ECG-CON and EDA-CON are not delayed (see Table 3).

For the smoothing of the initial prediction results, we adopt two Gaussian smoothing methods: one with the moving window of fixed length (120 frames), the other with windows of variable length. In the latter smoothing method, suppose the current frame number is T, then the length of the Gaussian window is set as T, and the smoothed value is calculated as the weighted average of the values from frame 1 to frame T as $x_T = (\sum_{i=1}^{T} w_i x_i)/T$, where the closer to the current frame, the higher the weight w_i is. However, if T is very high, the calculation of the weighted average will be time consuming. Therefore, in the implementation process, we adopt the non-uniform sampling strategy: more sampling points are used for the recent history, and less sampling points for the far history.

Figure 4 shows the arousal curves of a segment of video, including the ground-truth labels, the initial prediction results, as well as the smoothed results with the fixed length moving Gaussian window (Smooth-Fixed) or variable length Gaussian window (Smooth-Variable). We can see that with Gaussian smoothing, the curve becomes smoother while still capturing the ground-truth trends, and comparing with the results from Smooth-Fixed, the smoothed results from Smooth-Variable are closer to the original predictions.

In Figure 4, we can also observe that Smooth-Variable causes a little delay on the initial predictions. As the offsets have been added on the Audio-CON, LGBP-TOP, LPQ-TOP, and geometric features at the beginning, this further delay may reduce the CCs and CCCs between the smoothed predictions and the ground-truth labels. Therefore, for these

features, Gaussian smoothing with fixed length moving window is performed, while for the ECG and EDA features, Gaussian smoothing with variable length windows is adopted.

Finally, the smoothed predictions are input into a second layer of DBLSTM-RNN model for the final prediction of an affective dimension.

4. EXPERIMENTS AND ANALYSIS

In the training of the DBLSTM-RNN models, we adopt the CURRENNT toolbox [37]. To overcome the possible over-fitting problem in training the DBLSTM models, we split the development set into two subsets: sequences from the first 6 subjects are used as a *sub-dev-1* set to adjust the parameters, and those from the remaining 3 subjects are used as a *sub-dev-2* set to verify the performance. Parameters that get good and close performances on both sets are used to train the DBLSTM models.

We should notice that since the ECG and EDA signals of 1 video clip in the training set are not good, when training the models using the ECG-CON-FS and EDA-CON-FS features, we use the features from the other 8 video clips.

The prediction performance is reported in terms of root-mean-square-error (RMSE), CC and CCC. In our experiments, these measurements are calculated by averaging over all the sequences in the development set or test set.

4.1 Single-modal Affect Prediction

Single-modal prediction results on the development set are shown in Table 4. One can see that: 1) for arousal prediction, the Audio-CON-FS feature obtains very promising results, with CCC reaching 0.800, followed by LPQ-TOP-FS with CCC up to 0.587. While the newly extracted ECG-CON-FS and EDA-CON-FS features do not improve the prediction performance. 2) for valence, the Geo-FS feature set obtains the highest CCC as 0.441, followed by the Audio-CON-FS features. The LGBP-TOP feature set also performs well in predicting valence. 3) in both arousal and valence prediction, the Audio-CON-FS feature set obtains a much better performance than the baseline audio features, showing that the proposed audio features in this paper are very promising in affect recognition.

4.2 Multi-modal Affect Prediction

In the multi-modal affective dimension prediction, for each dimension (arousal/valence), we rank the features which obtained good performances in the single-modal prediction task, and input them into the DBLSTM based multi-modal prediction framework, as shown in Figure 3. Prediction results of arousal on the development set, with different combinations of the top ranked features, are shown in Table 5, and those of valence are given in Table 6. One can notice that: 1) for arousal prediction, the best performance, with $CCC = 0.824$, is obtained with the feature combination B composed of Audio-CON-FS, LPQ-TOP-FS, and the baseline ECG features. While when adding the baseline EDA features (combination C) and LGBP-TOP features (combination D), the performance is not further improved. 2) for valence prediction, the highest CCC (0.688) is obtained with the feature combination I composed of all the six extracted feature sets (Geo-FS, Audio-CON-FS, LGBP-TOP, ECG-CON-FS, LPQ-TOP-FS, and the baseline EDA features).

To verify the generalization ability of the DBLSTM-RNN based multimodal affect recognition system, we list the prediction results on *sub-dev-1* (used to adjust the parameters) and *sub-dev-2* (used to test the performance) in Table 7. Due to space limitation, we only list those from the best feature combinations, i.e. combination B for arousal, and combination I for valence, respectively. One can see that for both arousal and valence, the performances on the two subsets are very close, showing that the proposed DBLSTM-RNN based fusion framework has a good generalization ability.

Finally, as the proposed ECG and EDA features are not extracted for 3 video clips of the test set, for the evaluation on the test set, we submit our prediction results from the best feature combinations with the baseline ECG and EDA features, i.e. combination B for arousal and combination $I1$ for valence. Table 8 lists our prediction results along with the baseline ones. One can see that the CCCs obtained with our proposed features and the DBLSTM-RNN based fusion framework are much higher than the baseline results, both on the development set and the test set.

Table 5: Development set - multi-modal prediction results of arousal

Modal	RMSE	CC	CCC
A:Audio-CON-FS+LPQ-TOP-FS	.097	.844	.814
B: A+ECG baseline	.098	.850	**.824**
C: B+EDA baseline	.097	.850	.820
D: C+LGBP-TOP	.104	.821	.792

Table 6: Development set - multi-modal prediction results of valence

Modal	RMSE	CC	CCC
E: Geo-FS+Audio-CON-FS	.091	.654	.590
F: E+LGBP-TOP	.085	.663	.621
G: F+ECG-CON-FS	.086	.721	.667
G1: F+LPQ-TOP-FS	.086	.662	.614
H: G+LPQ-TOP-FS	.091	.709	.665
H1:G1+ECG baseline	.095	.665	.620
I: H+EDA baseline	.088	.725	**.688**
I1: H1+EDA baseline	.088	.665	.621

Table 7: Prediction results on *sub-dev-1* and *sub-dev-2*

Set	Arousal(B)			Valence(I)		
	RMSE	CC	CCC	RMSE	CC	CCC
sub-dev-1	.099	.864	.834	.093	.666	.622
sub-dev-2	.104	.788	.770	.093	.741	.705

5. CONCLUSIONS

This paper presents our system design for AV^+EC 2015. Besides the baseline features, we proposed extra audio features as functionals on the LLDs obtained via the YAAFE toolbox, LPQ-TOP features from the video sequences as well as various ECG based and EDA based features. For feature selection, the concordance correlation coefficient (CCC), instead of the usual Pearson correlation coefficient (CC), has been used as objective function. In addition, offsets between the features and the arousal/valence labels were considered in both feature selection and modeling of the affective dimensions. For multimodal continuous affect predic-

Table 4: Single modal prediction results on the development set

Dimension	Measures	Audio baseline	Audio -CON-FS	LGBP -TOP	LPQ -TOP-FS	Geo -FS	ECG baseline	ECG -CON-FS	EDA baseline	EDA -CON-FS
Arousal	RMSE	.148	.099	.185	.148	.179	.240	.204	.233	.188
	CC	.529	.836	.399	.665	.354	.444	.384	.345	.322
	CCC	.387	**.800**	.226	.587	.173	.297	.284	.248	.202
Valence	RMSE	.116	.104	.105	.114	.094	.228	.165	.156	.129
	CC	.167	.529	.501	.399	.562	.371	.344	.329	.246
	CCC	.112	.398	.346	.285	**.441**	.238	.293	.231	.189

Table 8: Comparison with baseline results on the development set and test set

Set	Arousal						Valence					
	baseline			our results(B)			baseline			our results(I)		
	RMSE	CC	CCC	RMSE	CC	CCC	RMSE	CC	CCC	RMSE	CC	CCC
Dev.	.161	.559	.476	.098	.850	.824	.105	.548	.461	.088	.725	.688
Test.	.164	.354	.444	.121	.753	.747	.113	.490	.382	.104	.616	.609

tion, we proposed a DBLSTM-RNN based fusion architecture, in which the initial predictions from the single modalities via the DBLSTM-RNNs are firstly smoothed with Gaussian smoothing, then input into a second layer of DBLSTM-RNN for the final prediction of an affective state. Experimental results show that our proposed features and the DBLSTM-RNN based fusion framework obtain very promising results on both the development set and test set of $AV^{+}EC$ 2015.

6. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (grant 61273265) and the VUB Interdisciplinary Research Program through the EMO-App project.

7. REFERENCES

[1] F. Agrafioti, D. Hatzinakos, and A. K. Anderson. ECG pattern analysis for emotion detection. *IEEE Transactions on Affective Computing*, 3(1):102–115, 2012.

[2] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Multi-scale temporal modelling for dimensional emotion recognition in video. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18. ACM, 2014.

[3] T. Costa, D. Galati, and E. Rognoni. The hurst exponent of cardiac response to positive and negative emotional film stimuli using wavelet. *Autonomic Neuroscience*, 151(2):183–185, 2009.

[4] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proc. of ACM MM*, pages 835–838. ACM, 2013.

[5] Y. Fan, Y. Qian, F. Xie, and F. K. Soong. TTS synthesis with bidirectional LSTM based recurrent neural networks. In *Proc. Interspeech*, 2014.

[6] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory. Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks. In *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014.

[7] A. Graves, N. Jaitly, and A.-r. Mohamed. Hybrid speech recognition with deep bidirectional LSTM. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 273–278. IEEE, 2013.

[8] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6645–6649. IEEE, 2013.

[9] X. Guo. Study of emotion recognition based on electrocardiogram and RBF neural network. *Procedia Engineering*, 15:2408–2412, 2011.

[10] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan. Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 33–40. ACM, 2014.

[11] H.Gunes, M. A. Nicolaou, and M. Pantic. Continuous analysis of affect from voice and face. *Computer Analysis of Human Behaviour. Springer Verlag, London*, pages 255–291, 2011.

[12] A. Jan, H. Meng, Y. Gaus, F. Zhang, and S. Turabzadeh. Hilbert-Huang transform based physiological signals analysis for emotion recognition. In *Signal Processing and Information Technology (ISSPIT), 2009 IEEE International Symposium on.* IEEE, 2009.

[13] A. Jan, H. Meng, Y. Gaus, F. Zhang, and S. Turabzadeh. Automatic depression scale prediction using facial expression dynamics and regression. In *Proceedings of the 4th ACM international workshop on Audio/visual emotion challenge*, pages 73–80. ACM, 2014.

[14] J. Kim and E. Andre. Emotion recognition based on physiological changes in music listening. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2067–2083, 2008.

[15] S. D. Kreibig. Autonomic nervous system activity in emotion: a review. *Biological Psychology*, 84(3):394–421, 2010.

[16] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard. YAAFE, an easy to use and efficient audio feature extraction software. *Proceedings of Ismir Conference*, 2010.

[17] H. Meng, D. Huang, H. Wang, H. Yang, M. Al-Shuraifi, and Y. Wang. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 21–29. ACM, 2013.

[18] V. Mitra, E. Shriberg, M. McLaren, A. Kathol, C. Richey, D. Vergyri, and M. Graciarena. The SRI AVEC-2014 evaluation system. In *Proceedings of the 4th ACM international workshop on Audio/visual emotion challenge*, pages 93–101. ACM, 2014.

[19] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *Affective Computing, IEEE Transactions on*, 2(2):92–105, 2011.

[20] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM international conference on multimodal interaction*, pages 501–508. ACM, 2012.

[21] P. Prenter. *Splines and Variational Methods*. Wiley, New York, 1989.

[22] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994.

[23] 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 *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC), ACM MM*, Brisbane, Australia, October 2015.

[24] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proceedings of Face and Gestures 2013, 2nd IEEE International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, Shanghai, China, April 2013.

[25] H. Sak, A. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014.

[26] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. AVEC 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011.

[27] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. AVEC 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012.

[28] G. Valenza, P. Allegrini, A. Lanata, and E. P. Scilingo. Dominant Lyapunov exponent and approximate entropy in heart rate variability during emotional visual elicitation. *Frontiers in Neuroengineering*, 5(3), 2012.

[29] G. Valenza, L. Citi, A. Lanata, E. Scilingo, and R. Barbieri. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. *Nature-SCIENTIFIC REPORTS*, 2014.

[30] G. Valenza, A. Lanata, and E. Scilingo. The role of nonlinear dynamics in affective valence and arousal recognition. *Affective Computing, IEEE Transactions On*, 3(2):237–249, 2012.

[31] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014.

[32] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013.

[33] L. Van Der Maaten. Audio-visual emotion challenge 2012: a simple approach. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 473–476. ACM, 2012.

[34] D. Ververidis and C. Kotropoulos. Fast sequential floating forward selection applied to emotional speech features estimated on des and susas data collections. In *Proc. XIV European Signal Processing Conf*, 2006.

[35] M. W02llmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll. LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153âĂŞ163, 2013.

[36] N. Wang, E. Ambikairajah, B. Celler, and N. Lovell. Accelerometry based classification of gait patterns using empirical mode decomposition. In *Proc. of ICASSP*, pages 617–620. IEEE, 2008.

[37] F. Weninger, J. Bergmann, and B. Schuller. Introducing CURRENNT–the Munich open-source CUDA recurrent neural network toolkit. *Journal of Machine Learning Research*, 15, 2014.

[38] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll. Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments. *Computer Speech and Language*, 28(4):888–902, 2014.

[39] 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 *INTERSPEECH*, pages 597–600, 2008.

[40] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *Selected Topics in Signal Processing, IEEE Journal of*, 4(5):867–881, 2010.