Multi-modal Dimensional Emotion Recognition using Recurrent Neural Networks

Shizhe Chen
Multimedia Computing Laboratory
School of Information
Renmin University of China
cszhe1@ruc.edu.cn

Qin Jin*
Multimedia Computing Laboratory
School of Information
Renmin University of China
qjin@ruc.edu.cn

ABSTRACT
Emotion recognition has been an active research area with both wide applications and big challenges. This paper presents our effort for the Audio/Visual Emotion Challenge (AVEC2015), whose goal is to explore utilizing audio, visual and physiological signals to continuously predict the value of the emotion dimensions (arousal and valence). Our system applies the Recurrent Neural Networks (RNN) to model temporal information. We explore various aspects to improve the prediction performance including: the dominant modalities for arousal and valence prediction, duration of features, novel loss functions, directions of Long Short Term Memory (LSTM), multi-task learning, different structures for early feature fusion and late fusion. Best settings are chosen according to the performance on the development set. Competitive experimental results compared with the baseline show the effectiveness of the proposed methods.

Categories and Subject Descriptors
I.5.4 [Pattern Recognition Applications]: *computer vision, signal processing*; J.4 [Social and Behavioral Sciences]: *Psychology, sociology.*

General Terms
Theory

Keywords
Affective Computing, Emotion Recognition, Recurrent Neural Network

1. INTRODUCTION
Automatic emotion recognition has been an active research area in past years, which is of great interest for human-computer interaction. It has wide application areas ranging from computer tutoring [1] to mental health diagnoses [2].

There are three major emotion computing models according to theories in psychology research [3]: discrete theory, dimensional theory, and appraisal theory. Discrete theory describes an emotion state as discrete labels such as "sad", "happy" etc. It is intuitive and simple but cannot express complex affective states. Dimensional theory considers an emotion state as a point in a continuous space. Hence, dimensional theory can model subtle, complicated, and continuous affective behavior. Typically an emotion state is covered by three dimensions: arousal (indicating the level of affective activation), valence (a measure of pleasure) and dominance (a measure of power or control). However, it is arguable that three dimensions are enough for describing so many emotion states. In addition, getting high quality dimensional labels is difficult. Appraisal theory attempts to detail the mental processes underlying the elicitation of emotions. It views an emotion state as a set of stimulus evaluation checks. But it is still an open research problem on how to use it. Most of the past research has focused on discrete emotion recognition [4]. However, more recently, affective computing researches are shifting towards dimensional emotion analysis [5-7] for better understanding of human emotions.

A broad range of modalities can express emotion information. The common cues include speech [4], text [8], facial expression [9], gesture [10], head movement, body movement/posture [11], and so on. Numerous findings in psychophysiology also suggest that there exists a correlation between bio-signals and affective states [12]. Combining different modalities seems to always improve emotion recognition performance. Most past works have fusion models on speech and text [13], audio and visual [14]. Results on fusion of audio, visual and physiological signals have not been well established.

The major issues for fusion of different modalities are: (i) when to integrate the modalities (i.e., at what abstraction level), and (ii) how to integrate the modalities (i.e., fusion methods) [15]. Decision level fusion assumes independence of different modalities. Feature level fusion assumes a strict time synchrony between the modalities and tends not to generalize well when the modalities substantially differ in temporal characteristics. Therefore, synchronization among different features becomes very important. The common fusion method is to ensemble multiple models such as random forest, AdaBoost, gradient boost regression tree and so on.

For dimensional emotion recognition, temporal information is very useful because the target dimensional values are continuous and have short time gap between two adjacent predictions. In this paper, we use temporal models, recurrent neural networks, to predict continuous dimensional values and explore many variations to improve performance. The major contributions investigated in this study can be listed as follows: (i) we use more powerful recurrent neural network models such as LSTM to capture contextual information (ii) we investigate features from audio, visual and physiology modalities, and find the dominant

*Corresponding Author

AVEC'15, October 26 2015, Brisbane, Australia
© 2015 ACM. ISBN 978-1-4503-3743-4/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2808196.2811638.

feature or feature combination for each dimension prediction (iii) we explore different LSTM structures and more effective loss functions to optimize the prediction performance. The performance on the development set is significantly improved over the original baseline, which shows the effectiveness of our methods.

The paper is structured as follows. Section 2 introduces related works in dimensional emotion recognition. Section 3 describes the dataset used in AVEC2015 challenge. Section 4 presents the proposed approaches and section 5 describes our extensive experimental results. Finally, section 6 concludes the paper.

2. RELATED WORK

In AVEC2014 challenge, the second place winner of the fully continuous sub-challenge [7] emphasizes temporal dynamic information using Deep Belief Network (DBN) with temporal pooling and multimodal-temporal fusion at the decision level, which demonstrates that high level temporal fusion can improve performance. The work also shows that the time delay between recordings and labels should be carefully processed.

Long Short Term Memory Recurrent Neural Networks (LSTM-RNN), one of the state-of-art sequence modeling techniques, has also been applied in dimensional emotion recognition. Wöllmer et al. [16] presented a fully automatic audiovisual recognition approach based on LSTM-RNN modeling of word-level audio and visual features. Compared to other models such as Support Vector Machine (SVM) and Conditional Random Field (CRF), LSTM achieved a higher prediction quality due to its capability of modelling long range time dependencies and decreasing the time delay. In AVEC2015 challenge, our experiments also demonstrate that LSTM outperforms traditional non-temporal models, so we focus on our investigation using the LSTM-RNN in this paper.

Meanwhile, other techniques based on deep neural networks (DNN) have been successfully applied in extracting emotional related features. For speech emotion recognition, work in [17] adds gender information to train auto-encoders and extracts the hidden layer as audio features and improves the unweighted emotion recognition accuracy. In work [18], convolutional neural networks (CNN) are applied in speech emotion task with novel loss functions to extract features. For facial expression recognition, the best result in the Emotion Challenge in the Wild 2013 is achieved by using CNN [19]. It has been proved that DNN can generate more powerful features.

Multi-task learning may also improve dimensional emotion prediction performance due to the correlation between arousal and valence. In work [20], two types of multi-task learning are introduced: one by learning each rater's individual track and the other by learning both dimensions simultaneously. Although it did not help for the audio feature based system, it improved the visual feature based system performance significantly.

Modalities fusion is another important issue in emotion recognition. Recent works [21] have studied synchronization between multimodal cues to support feature-level fusion and report greater overall accuracy compared to decision-level fusion.

Previous studies [20] also provide some common insights: 1) It's highly agreed that arousal is easier learned than valence. One reason is that the perception of arousal is more universal than is the perception of valence. 2) Audio modality is suitable for arousal prediction, but much less accurate for valence prediction. 3) Valence appears to be more stable than arousal using facial expression modality. Bio-signals are also good for valence assessment.

3. DATASET

The AVEC2015 challenge is evaluated on a subset of the RECOLA dataset [22], a multimodal corpus of remote and collaborative affective interactions. Subjects were divided into pairs to resolve a collaborative task ("winter survival task"), and data from different modalities such as audio, video, electro-cardiogram (ECG) and electro-dermal activity (EDA) were collected for each participant during their interactions. The data provided in the challenge consist of the first five minutes of the 27 speakers, which are unsegmented, non-prototypical and non-preselected. Emotional dimensions (arousal and valence) are annotated by 6 French speakers in scale [-1, 1] for every 40ms. Gold standard is calculated using a specific normalization technique as reported in [20]. Finally, the dataset is equally split into three partitions: train, development and test, with each partition containing 9 different speakers. .

The concordance correlation coefficient (CCC) works as the evaluation metric for this challenge, which is defined as:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2} \quad (1)$$

where μ_x and μ_y are the means for the two variables x and y, and σ_x^2 and σ_y^2 are the corresponding variances. ρ is the correlation coefficient between the two variables. Since it combines the Pearson Correlation Coefficient (CC) and the mean square error (mse), CCC is a much more reliable measurement for regression problems.

4. APPROACHES

4.1 Audio Features

Table 1: Short-time Low-level Acoustic Features

FEATURES	DESCRIPTION
Loudness+Delta	The loudness as the normalized intensity raised to a power of 0.3
F0final+Delta	The smoothed fundamental frequency contour
F0finEnv+Delta	The envelope of the smoothed fundamental frequency contour
jitterLocal+Delta	The local (frame-to-frame) Jitter (pitch period length deviations)
jitterDDP+Delta	The differential frame-to-frame Jitter (the 'Jitter of the Jitter')
shimmerLocal +Delta	The local (frame-to-frame) Shimmer (amplitude deviations between pitch periods)
Voicing final +Delta	The voicing probability of the final fundamental frequency candidate.
MFCC-related	MFCCs (15)+logMelFreqBand(8)

The AVEC2015 challenge provided baseline audio features are extracted by calculating the statistical functions over low-level descriptors in fixed window length of 3s with shift 40ms. We consider it as long-time audio feature. In order to compare with the long-time audio features in the baseline paper [22], we utilize the OpenSMILE toolkit [23] to extract short-time low-level frame

features [24]. The configuration file is modified according to the configuration file "emobase2010.conf" based on the Interspeech 2010 Paralinguistic Challenge [25]. Features are listed in Table 1. All the features are extracted with 40ms frame window and 40ms shift. Baseline long-time audio features have 102 dimensions, while short-time features consist of 76 dimensions.

4.2 Visual and Physiological Features

The Visual and physiological features used in this paper are the same as the provided baseline features used in [22]. The AVEC2015 challenge provides two sets of visual features extracted from facial expressions: appearance-based feature and geometric-based feature. The appearance-based feature is computed by using Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) and applied PCA to compress to 84 dimensions. The geometric-based feature is computed from 49 facial landmarks to generate 316-dimensional features. The features for frames without faces are padded with zeros. We consider both appearance-based and geometric-based visual features as short-time features. We consider ECG and EDA features as long-time features that are extracted with a sliding window of 4s in duration and 40ms shift. In total, there are 54 ECG features and 62 EDA features.

4.3 Temporal Model: LSTM

Long short term memory (LSTM) architecture [26] is the state-of-art model for sequence analysis since it uses memory cells to store information so that it can exploit long range dependencies in the data.

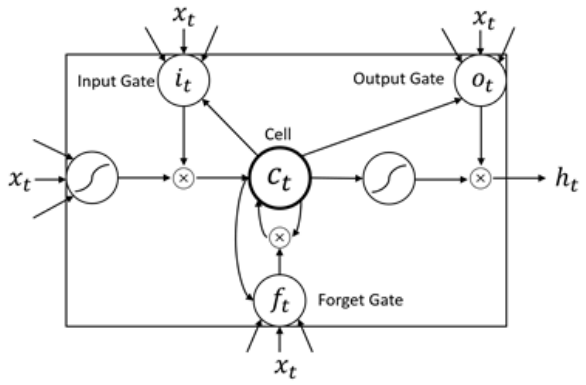

Figure 1: Long Short Term Memory Cell

In this paper, we use the LSTM version in [27]. Figure 1 illustrates a single memory cell. The functions of hidden cells and gates are defined as follows.

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
c_t &= f_t c_{t-1} + i_t \cdot tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \\
h_t &= o_t \cdot tanh(c_t)
\end{aligned}
\tag{2}
$$

where *i*, *f*, *o* and *c* refers to the input gate, forget gate, output gate, and cell input activation vectors respectively. $\sigma(\cdot)$ is the sigmoid function and $tanh(\cdot)$ is the tangent function. In [28], various architectures of recurrent neural network are explored and LSTM with the large forget bias achieved best performance on many tasks empirically. So we initialize our forget gate bias to a large value 2.

In addition to using LSTM memory cell as hidden units, we also employ bidirectional LSTM. A bidirectional RNN [29] makes use of all the past and all the future inputs on a single unit, which makes it ideal for processing data with de-synchronization between inputs and targets. Bidirectional property is achieved by processing the data sequences forwards and backwards in two separate hidden layers. The outputs from both hidden layers are then connected to the same output layer which fuses them. The combination of the concept of bidirectional RNN and LSTM leads to BLSTM.

4.4 Loss Function

We explore two types of loss functions in this paper: single point loss and overall correlation loss.

Single point loss is defined as in the following formula:

$$\text{Loss} = \frac{1}{T}\sum_{t=1}^{T} l(y_t, \hat{y}_t) \tag{3}$$

where *T* is the number of time steps in the sequence, y_t and $\hat{y}_t$ are the target emotion dimension value and the predicted value respectively at step t, l is a function to compute the distance between the target and the prediction. The most common distance is computed as mean square error (mse) or mean absolute error (mae). Similar with the support vector regression (SVR), we also compute the tube error, which does not count the absolute distance in the tube. This may have emphasis on the "super vector" data and make the predictions smoother. The formula is as follows:

$$
\begin{aligned}
l_{mse}(y_t, \hat{y}_t) &= (y_t - \hat{y}_t)^2 \\
l_{mae}(y_t, \hat{y}_t) &= |y_t - \hat{y}_t| \\
l_{tube}(y_t, \hat{y}_t) &= \begin{cases} 0, & if\ |y_t - \hat{y}_t| < \text{tube} \\ |y_t - \hat{y}_t| - \text{tube}, & otherwise \end{cases}
\end{aligned}
\tag{4}
$$

Overall correlation loss is computed using the overall characteristics between two sequences. The correlation loss function we use in this paper includes: Covariance (cov) and Pearson correlation covariance (CC).

Assuming target sequence $\mathrm{y} = (y_1, y_2, \ldots, y_T$ and the predicted sequence $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T)$. The covariance of these two sequences is defined as:

$$\text{cov}(\mathrm{y}, \hat{y}) = \frac{1}{T}\sum_{i=1}^{T}(y_i - \bar{y})(\hat{y}_i - \frac{1}{T}\sum_{j=1}^{T}\hat{y}_j) \tag{5}$$

where $\bar{y} = \frac{1}{T}\sum_{j=1}^{T} y_j$, so the derivative of $\hat{y}$ is as follows:

$$\frac{\partial \text{cov}(\mathrm{y}, \hat{y})}{\partial \hat{y}} = \frac{1}{T}(y - \bar{y}) \tag{6}$$

We can see from formulation (6) that the backpropagation loss has nothing to do with the prediction output. The intuitive understanding of this gradient is that it only helps to push the shape of prediction to be similar with the shape of target sequence without considering the real value of the prediction output. So distance constraint is necessary for overall correlation loss. We explore different loss functions including: 1) we use mse as distance constraint and use (mse-cov) as the overall correlation loss. 2) we use mse/(CC+1) as the overall correlation loss since Pearson correlation covariance (CC) which ranges between -1 and 1 can be considered as a global correlation among covariance dimensions. 3) concordance correlation covariance (CCC) can also be considered as an overall loss function with distance constraint.

4.5 Feature Fusion Structure

The most common way for feature fusion is the simple concatenation of features, but this suffers from asynchronization of features with different durations. One possible way to solve this problem is to input features with different duration to different layers of the network. It might relax the compatibility problem in time for different features. For example, Figure 2 illustrates such a LSTM structure for this purpose. The first hidden layer takes the short-time audio features as input. The second layer takes the longer-time visual features concatenated with the output of the first layer as input. The last hidden layer takes the longest-time ecg features concatenated with the output of the second layer as input. It is generally reported that the higher hidden layers capture more abstract and more global information from the source; we expect this new proposed feature fusion structure can alleviate the asynchronization problem.

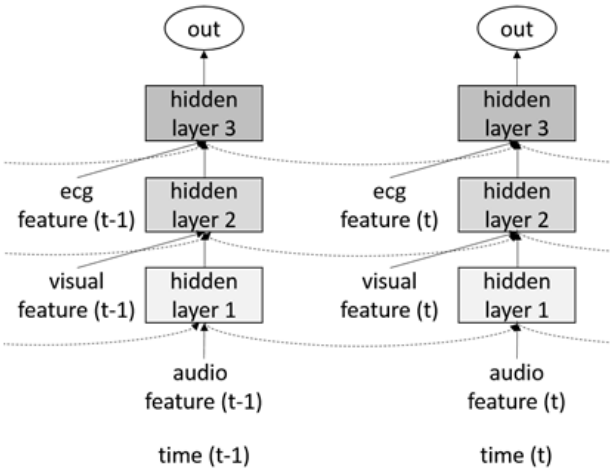

Figure 2: New feature fusion network structure

4.6 Smoothing

The predicted target from the temporal model can be noisy so smoothing is required to get better performance. Exponential smoothing is a rule of thumb technique for smoothing time series data [30]. It can take into account all the past data. The form of exponential smoothing is given by the formula:

$$s_t = \alpha \cdot x_t + (1-\alpha) \cdot s_{t-1} \quad (7)$$

where x_t and s_t are the sequence value and the smoothed value at time t respectively. α ($0 < \alpha < 1$) is the smoothing factor. Larger α has less smoothing effect and gives greater weight to local values, while smaller α has better smoothing effect. The optimization of α is done by grid search. The initial value of s_1 is x_1. In our experiments, we find that if LSTM is forward, applying smoothing backwardly achieves slightly better performance, while if LSTM is backward, applying smoothing forwardly is better.

5. EXPERIMENTS

All the features in following experiments are normalized in each speaker set using z-score. The gradient descend method is rmsprop [31] with nesterov momentum [32] of 0.9 and learning rate of 1e-4. Early stopping strategy is used to avoid overfitting. Training stops if there is no improvement after 50 epochs. The Neural network in this paper is implemented using Theanets [33], a wrapped library based on Theano [28].

The structure of LSTM is presented as a list in this paper. For example, [618,160,120,1] indicates that the network contains an input layer with 618-dimentional input feature, two hidden layers with hidden units of 160 and 120 respectively, and an output layer with one node corresponding to the predicted emotion dimension value.

5.1 Dominant Features

In this set of experiments, we aim to investigate best features for each emotion dimension prediction. First, we apply a two-hidden layer LSTM to single modality features in the baseline to find the dominant feature for predicting each emotion dimension. Due to time and resource limitation in this challenge, the size of the two hidden layers is only optimized among (80,60), (160,120), (200,160) three setups respectively for each modality and mse loss function is applied.

Table 2 shows the prediction performance using each single modality features on the development set. As we can see from the table that audio features achieve the best performance on the arousal dimension prediction. For valence prediction, visual features perform the best. Physiological features based emotion dimension prediction is however relatively low. Figure 3 presents an example of the arousal dimension prediction based on the eda feature, which indicates that prediction with CCC below 0.1 is not reliable. So we only consider features with CCC above 0.1 for fusion later. As expected, arousal prediction is an easier task than valence prediction.

Table 2: Performance of baseline single modality feature on the development set

	Arousal			Valence		
modality	rmse	CC	CCC	rmse	CC	CCC
audio	**0.133**	**0.730**	**0.703**	0.118	0.369	0.312
video_app	0.168	0.550	0.533	0.114	0.430	0.354
video_geo	0.182	0.470	0.417	**0.111**	**0.523**	**0.504**
ecg	0.190	0.343	0.221	0.119	0.303	0.159
eda	0.195	0.140	0.065	0.117	0.348	0.236

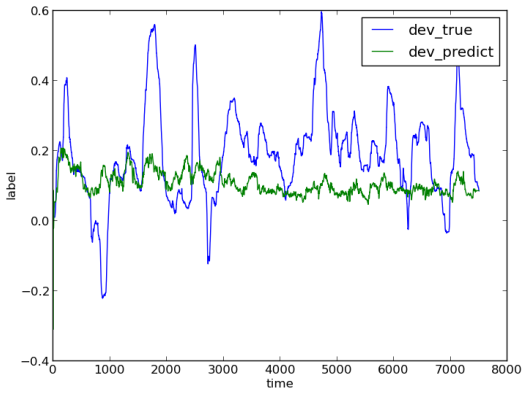

Figure 3: An arousal prediction example on the development data based on eda feature

5.2 Long-time vs. Short-time Features

The original provided audio feature is from applying statistical functions in 3s analysis window. We consider it as long-time audio feature. Since LSTM is good at memorizing, we input short-time audio features with 40ms analysis window to LSTM. The

network size is optimized with two hidden layers of size (160, 120). As shown in Table 3, short-time audio feature outperforms long-time feature for arousal prediction while long-time audio feature works better on valence prediction, which is consistent with the knowledge that valence is more related to longer time period. So in our following experiments, we only use short-time audio feature for arousal prediction and long-time audio feature for valence prediction.

Table 3: Performance based on long-time and short-time audio features

Target	Feature	rmse	CC	CCC
Arousal	long-time audio	0.133	0.730	0.703
	short-time audio	***0.122***	***0.78***	***0.762***
Valence	**long-time audio**	**0.118**	**0.369**	**0.312**
	short-time audio	*0.119*	*0.32*	*0.224*

Figure 4 shows an arousal prediction example based on long-time and short-time audio feature. The red curve shows the true arousal value. The green and blue curves show the predicted arousal value based on the long-time audio feature and the short-time audio feature respectively. From Figure 4, we can see that curve fitting is better in peaks and valleys with short-time audio feature. This may relate to the fact that long-time features may lose some detailed local information compared with short-time features.

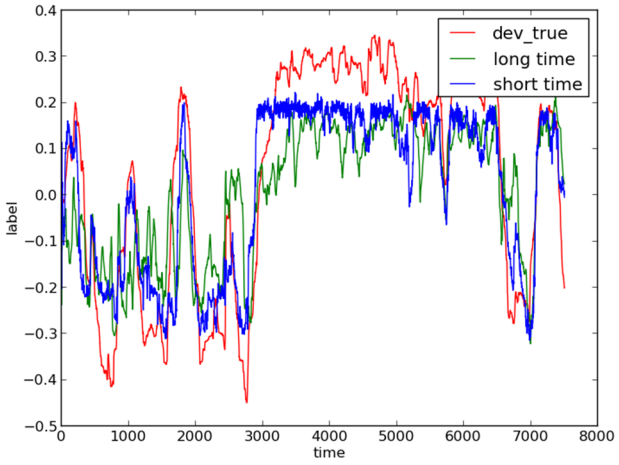

Figure 4: An arousal prediction example based on the long-time audio feature and the short-time audio feature

5.3 Comparison of Loss Functions

Table 4 and Table 5 compare the prediction performance for arousal and valence dimensions respectively using different loss functions. For single point loss functions, mse achieves the best performance in general. mae and tube error seem unstable for different modality features. Overall covariance loss function with distance constraints achieves much better performance especially with loss functions (-CCC) and (mse-cov), which improve the single modality performance significantly. Loss function (mse-cov) tends to work better for arousal prediction and loss function (-CCC) tends to work better for valence prediction.

5.4 Bidirectional LSTM

We compare the directions of LSTM in this section. BLSTM is to combine a forward LSTM and a backward LSTM in an elegant way, of which one hidden units can get input from the past and the future. So for BLSTM we double the size of LSTM. We choose the best single modality with the corresponding best loss function to compare the performance of BLSTM and LSTM as shown in Table 6. Although BLSTM achieves slightly better performance than LSTM, it is much more time consuming and prone to overfitting. We therefore consider that LSTM might be good enough for this task.

Table 4: CCC performance comparison with different loss functions for arousal prediction

	Single Point Loss			Overall Covariance Loss		
	mse	mae	tube0.05	-ccc	mse-cov	mse/(CC+1)
short-time audio	0.762	0.802	0.756	**0.807**	0.761	0.765
video_app	0.533	0.258	0.511	0.515	**0.571**	0.476
video_geo	0.417	0.340	0.394	0.423	**0.471**	0.354
ecg	0.221	0.240	0.235	**0.333**	0.329	0.208
Average	0.483	0.410	0.474	0.520	**0.533**	0.451

Table 5: CCC performance comparison with different loss functions for valence prediction

	Single Point Loss			Overall Covariance Loss		
	mse	mae	tube0.05	-ccc	mse-cov	mse/(CC+1)
long-time audio	0.312	0.24	0.221	0.334	**0.443**	0.435
video_app	0.354	0.409	0.415	**0.526**	0.496	0.433
video_geo	0.504	0.517	0.468	**0.557**	0.530	0.525
ecg	0.159	0.114	0.147	**0.314**	0.252	0.229
eda	0.236	0.131	0.174	**0.315**	0.262	0.140
Average	0.313	0.282	0.258	**0.409**	0.397	0.352

Table 6: Comparison of BLSTM and LSTM

	Short-time audio, arousal			**Video_geo, valence**		
	RMSE	CC	CCC	RMSE	CC	CCC
LSTM	0.115	0.810	0.807	0.116	0.564	0.557
BLSTM	0.116	0.816	0.812	0.121	0.573	0.564

5.5 Multi-task learning

We explore utilizing multi-task learning scheme to learn arousal and valence dimension prediction together using different features. First, we explore multi-task learning using the best single modalities (short-time or long-time audio features and video_geo feature). As shown in Table 7, we compare the performance of our best single modality system with and without multi-task

learning. The upper part refers to our baseline single modality system performance while the lower part (bolded) refers to the performance with multi-task learning. We can see from the table that multi-task learning improve the arousal prediction very marginally. For valence prediction, multi-task learning improves the performance based on video_geo feature but decreases the performance based on long-time audio features. We suspect the reason might be that there is larger gap between arousal prediction and valence prediction based on audio features, while arousal and valence predictions based on video_geo features are more balanced. We also investigate multi-task learning with all modality features combined/concatenated and with bigger network size; however we could not achieve better performance.

Table 7: Performance comparison with and without multi-task learning

	Arousal			Valence		
	RMSE	CC	CCC	RMSE	CC	CCC
Audio	0.122	0.780	0.762	0.118	0.369	0.312
Video_geo	0.182	0.470	0.417	0.111	0.523	0.504
Audio	**0.134**	**0.771**	**0.765**	**0.119**	**0.379**	**0.235**
Video_geo	**0.195**	**0.426**	**0.418**	**0.104**	**0.622**	**0.618**

5.6 Early Fusion

Early fusion or feature fusion in this work is implemented either by feature concatenation or new feature fusion structure as described in section 4.5. For experiments in this subsection, we fix the LSTM structure with 3 hidden layers of size (200,200,200). All modality features are applied for valence prediction. For arousal prediction we drop the eda feature as its CCC performance is below 0.1. Results are shown in Table 8. From Table 8 we can see that new feature structure could not outperform the simple concatenation feature fusion on all three modalities.

Table 8: Comparison of new feature fusion structure and feature concatenation using three modalities

Target	Architecture	rmse	CC	CCC
Arousal	concatenation	0.149	0.802	0.777
	new structure	0.141	0.789	**0.779**
Valence	concatenation	0.101	0.654	**0.652**
	new structure	0.107	0.613	0.610

Table 9: Comparison of new feature fusion structure and feature concatenation using two modalities

feature	architecture	CCC
Audio & Video (video_app+video_geo)	Concatenation	0.609
	New structure	0.536
Video (video_app+video_geo) & BioSignal (ecg+eda)	Concatenation	**0.588**
	New structure	**0.611**
Audio & BioSignal (ecg+eda)	Concatenation	0.492
	New structure	0.451

We also compare two fusion strategies based on two modalities. The network structure consists of two hidden layers with 200 units for each layer. Results are presented in Table 9. We can see that new feature fusion structure improves the video and bio-signal modalities feature fusion over simple feature concatenation. However, it cannot outperform simple concatenation when audio modality is used.

5.7 Late Fusion

Late fusion/decision fusion is a linear regression of the different predictions from multiple subsystems for each emotional dimension prediction. We further split the development set into two parts, one part with the first 8 sequences for learning fusion weights (dev-dev) and the last sequence for testing fusion performance (dev-test). We apply smoothing before fusion. For arousal prediction, we fuse the predictions of the top loss function for each modality, which are: (long-time audio, mse), (short-time audio, mae), (short-time audio, mse-cov), (video_app, mse), (video_app, mse-cov), (video_geo, -ccc), (video_geo, mse-cov), (ecg, -ccc), (ecg, mse-cov). We conduct similar late fusion for valence prediction as well. As shown in Table 10, late fusion improves the final predictions (0.841 vs 0.768 for arousal, 0.669 vs 0.597 for valence).

Table 10: Late fusion Performance

	Arousal	Valence
best baseline on dev w/o fusion	0.768	0.597
with late fusion on dev-dev	0.872	0.717
with late fusion on dev-test	0.770	0.600
with late fusion on entire dev	0.841	0.669

5.8 Smoothing Effects

We apply exponential smoothing (as described in subsection 4.6) on the raw prediction from LSTM. Table 11 presents some examples to show the effectiveness of smoothing. Smoothing can bring some additional small improvement.

Table 11: Performance comparison w/o smoothing

target	feature	Alpha	Original	Smoothed
Arousal	short-time audio	0.1	0.8073	0.8101
Valence	video_geo	0.7	0.5570	0.5571

5.9 Best Submitted Run

Table 12 presents the best arousal prediction and the best valence prediction from our 5 submitted runs. The best arousal prediction system is from using late fusion of all modalities as described in section 5.7 and further late fused with two early fusion systems (concatenation and new structure). The best valence prediction is from using concatenation early feature fusion as described in section 5.6. The prediction performance is significantly improved over the provided baseline performance in [21].

Table 12: The best submission results

	Arousal			Valence		
	rmse	CC	CCC	rmse	CC	CCC
train	0.074	0.974	0.928	0.043	0.973	0.952
dev	0.096	0.867	0.860	0.101	0.654	0.652
test	0.121	0.746	0.739	0.111	0.590	0.567

6. CONCLUSIONS

This paper presents our approach to model dimensional emotions using recurrent neural networks. We explore different aspects for improving prediction performance, including various modality features, duration of features, novel loss functions, bidirectional networks, early fusion and late fusions. Some findings from extensive experiments are listed as follows:

(1) Audio feature alone performs the best to predict the arousal dimension while visual feature is more suitable for valence dimension prediction.

(2) Loss functions related to overall covariance with distance constraint are good for dimensional emotion predictions.

(3) Single direction of LSTM is good enough for dimensional emotion regression in this task.

(4) Early feature fusion from different modalities helps valence dimension prediction. Late fusion always helps.

(5) Inputting features that are not compatible in time into different layers of network can help with feature asynchronization.

Our future work will focus on improving predictions by extracting more powerful modality features and improving RNN structures.

7. ACKNOWLEGEMENT

This work was partially supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNLQ01), the Beijing Natural Science Foundation (No. 4142029), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

8. REFERENCES

[1] D. Litman and K. Forbes, Recognizing emotions from student speech in tutoring dialogues. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003.

[2] D.J. France, R.G. Shiavi, S. Silverman, M. Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans. Biomedical Eng.* 47(7) (2000) pp. 829-837.

[3] Stacy Marsella. Computationally Modeling Human Emotion. *Communications of the ACM*, Vol. 57 No. 12, Pages 56-67, 2015.

[4] M. Ayadi, M. Kamel, F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(2011) 572-587.

[5] D. Wu, T. D. Parsons, E. Mower, S. Narayanan. Speech emotion estimation in 3D space. *Multimedia and Expo (ICME)*, 2010.

[6] M. Wollmer, M. Kaiser, F. Eyben, B. Schuller. LSTM Modelling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 2012.

[7] L. Chao, J. Tao, M. Yang, Y. Li, Z. Wen. Multi-scale temporal modeling for dimensional emotion recognition in video. proc. 4rd *ACM international workshop on Audio/Visual emotion challenge*, 2014.

[8] R. Socher, A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C.P. Potts. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. Conference on *Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[9] Yichuan Tang. Deep Learning with Linear Support Vector Machines. In Workshop on *Representational Learning*, ICML 2013, Atlanta, USA, 2013.

[10] S. Piana, A. Staglianò, F. Odone, A. Verri, A. Camurri. Real-time Automatic Emotion Recognition from Body Gestures. http://arxiv.org/pdf/1402.5047v1.pdf

[11] G. Castellano, S. D. Villalba, A. Camurri. Recognising Human Emotions from Body Movement and Gesture Dynamics. *Affective Computing and Intelligent Interaction*, 2007.

[12] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun. Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals (*Tech. Rep. 05.02*). Geneva, Switzerland: Computer Vision Group, Computing Science Center, University of Geneva, 2002.

[13] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, A. Vembu, R. Prasad. Emotion Recognition using Acoustic and Lexical Features. *INTERSPEECH*, 2012.

[14] Z. Zeng, M. Pantic, G. I. Rosiman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[15] H. Gunes, M. Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1): 68-99, 2010.

[16] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, G. Rigoll. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, Volume 31, Issue 2, Pages 153-163, February 2013.

[17] R. Xia, J. Deng, B. Schuller, Y. Liu. Modeling gender information for emotion recognition using Denoising autoencoder. 990-994, *ICASSP*, 2014.

[18] Z. Huang, M. Dong, Q. Mao, Y. Zhan. Speech Emotion Recognition Using CNN. Proceedings of the *ACM International Conference on Multimedia*, MM14, Orlando, FL, USA, 2014.

[19] S.E. Kahou, C. pal, X. Bouthillier, P.Froumenty, C. Gulcehre, R. Memisevic, P. Vincent, A. Courville, and Y. Bengio. Combining Modality Specific Deep Neural Networks for Emotion Recognition in Video. In Proceedings of the 15th *ACM International Conference on Multimodal Interaction* (ICMI'13) pp. 543-550, 2013.

[20] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.P. Thiran, T. Ebrahimi, D. Lalanne, B. Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 29 November 2014.

[21] C. Shan, S. Gong, and P.W. McOwan. Beyond facial expressions: Learning human emotion from body gestures. *British Machine Vision Conference*, Warwick, UK, 2007.

[22] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, M. Pantic. The AV+EC 2015 Multimodal Affect Recognition Challenge: Bridging Across Audio, Video, and Physiological Data, *AVEC Workshop*, 2015.

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

[24] Q. Jin, C. Li, S. Chen, H. Wu, Speech Emotion Recognition With Acoustic And Lexical Features, *ICASSP*, Brisbane, Australia, 2015.

[25] B. Schuller, A. Batliner, S. Steidl, D. Seppi. Recognizing Realistic Emotions and Affect in Speech: State of the Art and Lessons Leant from the First Challenge. *Speech Communication*, 53(10), pp. 1062-1087, 2011.

[26] S. Hochreiter, J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997, 9(8):1735-1780.

[27] Alex Graves. Generating Sequences with Recurrent Neural Networks. http://arxiv.org/pdf/1308.0850v5.pdf, 2013.

[28] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley and Y. Bengio. Theano: new features and speed improvements. *NIPS 2012 deep learning workshop*.

[29] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. IEEE *Transactions on Signal Processing*, 1997, 45(11): 2673—268.

[30] Robert Goodell Brown. Smoothing Forecasting and Prediction of Discrete Time Series. Englewood Cliffs, NJ: Prentice-Hall, 1963.

[31] T. Tieleman, and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 4, 2012.

[32] I. Sutskever, J. Martens, G. Dahl et al. On the importance of initialization and momentum in deep learning. Proceedings of *International Conference on Machine Learning*, 2013.

[33] Theanets: https://github.com/lmjohns3/theanets/