Prediction-based learning for continuous emotion recognition in speech

Jing Han, Zixing Zhang, Fabien Ringeval, Björn Schuller

PREDICTION-BASED LEARNING FOR CONTINUOUS EMOTION RECOGNITION IN SPEECH

Jing Han[1]*, Zixing Zhang*[1]*, Fabien Ringeval*[2]*, Björn Schuller*[1,3]

[1]Chair of Complex & Intelligent System, University of Passau, Passau, Germany
[2]Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, Grenoble, France
[3]Department of Computing, Imperial College London, London, UK
jing.han@uni-passau.de

ABSTRACT

In this paper, a prediction-based learning framework is proposed for a continuous prediction task of emotion recognition from speech, which is one of the key components of affective computing in multimedia. The main goal of this framework is to utmost exploit the individual advantages of different regression models cooperatively. To this end, we take two widely used regression models for example, i. e., support vector regression and bidirectional long short-term memory recurrent neural network. We concatenate the two models in a tandem structure by different ways, forming a united cascaded framework. The outputs predicted by the former model are combined together with the original features as the input of the following model for final predictions. The experimental results on a time- and value-continuous spontaneous emotion database (RECOLA) show that, the prediction-based learning framework significantly outperforms the individual models for both arousal and valence dimensions, and provides significantly better results in comparison to other state-of-the-art methodologies on this corpus.

Index Terms— Affective computing, hierarchical regression models, support vector regression, long short-term memory

1. INTRODUCTION

In recent years, increasing efforts in the affective computing community have been paid on the automatic and *continuous* emotion prediction of humans' spontaneous behaviour [1–3]. For such a regression problem, a variety of regression models have been proposed and investigated, such as Support Vector Regression (SVR) [4], Relevance Vector Regression (RVR) [5], Feedforward Neural Networks (FNNs) [6], Deep Belief Networks (DBNs) [7], and Recurrent Neural Networks (RNNs) [8]. To better choose the regressors, distinct advantages of each model have been compared in the literature. For instance, the work in [9] has compared the performance of SVR and Bidirectional Long Short-Term RNNs (BLSTM-RNNs) for the continuous prediction of arousal and valence on the Sensitive Artificial Listener database, and the results indicate that the latter performed better when using 15 acoustic Low-Level-Descriptors (LLDs) as the feature set. Whereas, an opposite conclusion has been drawn in the work of [10], where SVR was superior to the BLSTM-RNN on the same database when using a statistical feature set that applied functionals over a large ensemble of LLDs. Other results in the literature confirm this inconsistent performance observation between SVR and diverse neural networks like (B)LSTM-RNNs and Feedforward Neural Networks (FNNs) [11]. A reasonable explanation behind this could be that each prediction model has its pros and cons. For example, BLSTM-RNNs are highly sensitive to overfitting, but SVR cannot explicitly model contextual dependencies.

In order to integrate the advantages of different regression models, some ensemble-based approaches have been introduced to further improve the continuous emotion prediction. The research in [12] has presented a stacked system, where LSTM-RNNs are first employed to predict the emotions in multi-dimensional and multi-modality spaces. Then, a Multiple Kernel SVR (MK-SVR) is used to correlate the generated multiple predictions to make a final decision. This method takes the benefits of LSTM-RNNs for modelling contextual dependencies, and MK-SVR for modelling the non-linear correlations between inputs and outputs. Furthermore, another similar structure has been introduced in [13], where predictions obtained by 20 variants of DBN topology structures are then aggregated by a SVR model. However, due to the similarity of characteristics of different DBNs, the predictions can not provide many variations that could be mutually complemented and improve the system performance. Besides, 20 DBNs are rather complex to be applied in the real world.

To further efficiently aggregate the advantages of different models, we put forward a simply constructed method. Different from the methods used in [13] and [12], we feed the predictions of the model in the first stage together with the original raw features into the model in the second stage. In doing this, only one raw feature set and two recognition models are required. The underlying idea is also in line with the expectation in the human community that one's suggestion can often ameliorate the others' judgement. Moreover, the idea is also inspired by a tandem structure for Automatic Speech Recognition (ASR) [14], where the phoneme predicted by neural networks is considered as an additional attribute for a Gaussian Mixture Model (GMM). In the present paper, we develop an prediction-based learning framework for the regression task of emotion recognition in speech.

In the remainder of this paper, we first briefly introduce the related work in Section 2. We then introduce the prediction-based learning method to exploit the predictive ability of SVR and BLSTM-RNN in Section 3. Afterwards, in Section 4, we implement and evaluate the system on the RECOLA dataset. Finally, we conclude our work and give an outlook in Section 5.

2. RELATED WORK

Our prediction-based method is related to the *Output Associative Relevance Vector Machine* (OA-RVM) regression framework originally proposed in [15]. The OA-RVM framework attempts to incorporate the contextual relationships that exist within and between dif-

ferent affective dimensions and various multimodal feature spaces, by training a secondary RVM with an initial set of multi-dimensional output predictions (learnt using any prediction scheme) concatenated with the original input features spaces. Additionally, the RVM framework also attempts to capture the temporal dynamics by employing a sliding window that incorporates both past and future initial outputs into the new feature space. Results presented in [16] also indicate that the OA-RVM framework, is well suited to emotion recognition problem.

The OA-RVM systems, like our proposed method, take input features and output predictions into consideration to train a subsequent regression model to perform the final affective predictions. The strength of the OA-RVM framework is that it is underpinned by the RVM. However, the results in [16] indicate that the framework is not as successful as it is when using either an SVR or a simple linear regressor as the secondary model. Further, the OA-RVM is non-casual and requires careful tuning to find a suitable window size so as to efficiently combine the initial outputs. This takes considerable time and effort. The proposed prediction-based learning framework, however, is designed to work with *any* combinations of learning paradigms, which aims to use the initial set of predictions to help improve the accuracy of any subsequent model. Furthermore, our method is simple constructed. It combines the original input features and the initial predictions frame by frame, which holds a strong advantage over the OA-RVM especially in the real-time applications.

3. PREDICTION-BASED LEARNING

3.1. Overview

The prediction-based learning framework for emotion recognition is depicted in Fig. 1. Specifically, in the *training* phase, the first regression model ($Model_1$) provides an initial prediction $\hat{y}_t$ based on the original acoustic feature vector $\boldsymbol{x}_t$. Then, the initial prediction $\hat{y}_t$ is concatenated with original feature vector $\boldsymbol{x}_t$ as the input $[\boldsymbol{x}_t, \hat{y}_t]$ of the second regression model ($Model_2$) for a final emotion prediction. An alternative strategy to train the second model is to employ a 'pseudo' prediction $\tilde{y}_t$ of $model_1$, which is simulated by applying noise to the true label. In this case, the combined feature set $[\boldsymbol{x}_t, \tilde{y}_t]$ is considered as the input of $Model_2$. In the *evaluation* phase, nevertheless, only the true initial prediction $\hat{y}_t$ is jointed with original feature vector $\boldsymbol{x}_t$, i. e., $[\boldsymbol{x}_t, \hat{y}_t]$, is employed as the input of $Model_2$ for a final emotion prediction.

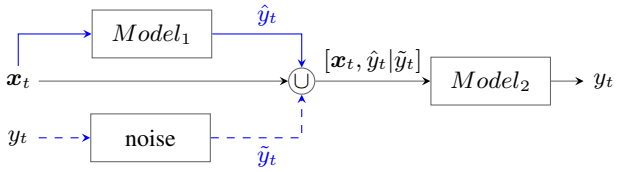

Fig. 1. Prediction-based learning framework

The motivation of using simulated prediction is that there is no need for us to know the type of $model_1$ beforehand. Therefore, it offers us a benefit that the $model_2$ can be simply and directly integrated with any pre-existing systems no matter what regression models they used (i. e., a plug and play). In this case, the prediction from the unknown model can be fused with the original feature set as the input of $Model_2$ for a final prediction in order to further improve the system performance.

In this preliminary study, we only take two widely used regression models, i.e., SVR and BLSTM-RNN, into consideration to evaluate the effectiveness of the proposed framework. In the rest of this section, the selected two models and the aforementioned two training schemes will be described in detail.

3.2. Regression Models

SVR is an extended version Support Vector Machine (SVM) to solve regression problems. It was first introduced in [4] and has become one of the most dominant methods in the context of machine learning, particularly for continuous emotion recognition [11].

Applying SVR for a regression task, the target is to optimise the generalisation bounds for regression by a loss function which is used to measure the cost of the errors of the prediction. Moreover, a predefined parameter C is set accordingly for different cases to balance emphasis on the errors and the generalisation performance. Usually, a (non-)linear kernel function is learned by the model to map the raw features into a higher mapped feature space. In our study, we use the SVR with a linear kernel function, as the features in the original feature space perform well for emotion prediction in our case. One of the most important advantages of SVR is that the global optimal solution is guaranteed owing to the characteristics of the convex optimisation function. Besides, SVR is learned by minimising an upper bound on the expected risk, as opposed to the neural networks trained by minimising the errors on all training data, which equips SVM a superior ability to generalise [17]. For more details about the SVR model, please refer to [4].

Another model utilised in our study is BLSTM-RNN which has been successfully applied to continuous emotion prediction [11] as well as for other regression tasks, such as non-linguistic vocalisations classification [2]. In general, it is composed of one input layer, one or multiple hidden layers, and one output layer [18]. The bidirectional hidden layers separately scan the input sequences in a forward and a backward order separately and connect to the same output layer. Compared with traditional RNNs, it introduces recurrently connected memory blocks to replace the network neurons in the hidden layers. Each block consists of a self-connected memory cell and three gate units, namely input, output, and forget gate. These three gates allow the network to learn when to write, read, or reset the value in the memory cell. Such a structure grants BLSTM-RNN to learn past and future context in both short and long range. For a more in-depth explanation of BLSTM-RNNs the reader is referred to [18].

3.3. Hierarchical System

In this study, we chose the models of SVR and BLSTM-RNN, the advantages and drawbacks of which are as follows

- the SVR model is more likely to achieve the globally optimal solution, but it is not context-sensitive [9];
- the BLSTM-RNN model is easily trapped in a local minimum which can be hardly avoided and has a risk of over-fitting [19], while it is good at capturing the correlation between the past and the future context [9].

Thus, by integrating the predictions from each of these two models, it is expected to overcome the limitations of each model whilst preserving the advantages of each. In other words, it aims to effectively make use of the advantages of the first model to complement the disadvantages of the second model.

As shown in Fig. 1, $Model_1$ and $Model_2$ in the framework could be either a SVR model or a BLSTM-RNN model, resulting in

four possible permutations, i.e., SVR-SVR, SVR-RNN, RNN-SVR, RNN-RNN. Particularly, it has to be noticed that the structure of RNN-RNN is regarded as a deep variation of neural networks. In addition, the structure of SVR-SVR is not considered since the training is done by solving a large margin separator, which will be likely the same if only adding its own prediction as one more attribute.

To train such a prediction-based learning system, the two models could be trained either *dependently* or *independently*, resulting in two training schemes as described below.

3.3.1. Training with True Predictions

When two models are trained *dependently*, $Model_2$ takes the predictive ability of $Model_1$ into account for training. The detailed procedure is given as follows:

- First, train $Model_1$ with the original feature set $\boldsymbol{x}_t$ to make an initial prediction $\hat{y}_t$.
- Then, train $Model_2$ with the combined feature set $[\boldsymbol{x}_t, \hat{y}_t]$ to learn the expected prediction y_t.

3.3.2. Training with Pseudo Predictions

$Model_1$ and $Model_2$ can also be trained *independently*, which means that training process of $model_2$ does not includes any information of $model_1$. In doing this, we use a pseudo prediction that is synthesised by adding noise to disturb the truth label y_t (gold standard or ground truth) to server as the additional attribute (denoted as $\tilde{y}_t$). Details about the training process are as follows:

- First, apply noise to the gold standard y_t to generate pseudo prediction $\tilde{y}_t$.
- Second, train $Model_2$ with the combined feature set $[\boldsymbol{x}_t, \tilde{y}_t]$ to learn the expected prediction y_t.

4. EXPERIMENTS

4.1. Data and Features

We evaluated the proposed method on the RECOLA corpus [20], which contains fully spontaneous and natural affective behaviours. It includes 46 multimodal (audio, video, and physiological data) recordings of French speaking participants involved in a dyadic collaborative task. Affective behaviour of the participants was evaluated by six different annotators (3 females), for the first five minutes of each recording, i.e., all annotators consistently annotated all recordings. Annotation was performed for arousal and valence separately. The obtained labels were then resampled to a constant $40\,ms$ frame rate, and averaged over all raters by considering inter-evaluator agreement, to provide a 'gold standard' [21]. In order to ensure speaker-independence in the experiments, the corpus was split into three partitions, by balancing the gender and the age of the subjects (cf. Table 1): training (16 subjects), validation (15 subjects), and testing (15 subjects). It is worth mentioning that, we used exactly the same partitions as in [3, 22] for comparison.

To extract acoustic features from the speech recordings, we used the openSMILE toolkit [23] to extract 13 LLDs, i.e., 12 Mel-Frequency Cepstral Coefficients, and 1 log energy with a frame window size of $25\,ms$ at a step size of $10\,ms$. Then, the arithmetic mean and the coefficient of variance were computed over the sequential LLDs within an analysed window of $8\,s$, resulting in 26 statistical features per window. The analysed window moves forward at the rate of $40\,ms$ such that it can match the granularity of the annotation. The total numbers of the segments in the train, development and test partitions were $120\,000$, $112\,500$, and $112\,500$, respectively.

Table 1. Three partitions of the RECOLA database.

#	train	development	test
female	10	9	8
male	6	6	7
French	11	11	11
Italian	3	2	3
German	2	1	1
Portuguese	0	1	0
age μ (σ)	22.3 (3.4)	21.6 (2.1)	21.2 (2.0)

4.2. Implementation and Evaluation

For the SVR models, we chose a linear kernel function with optimised C in [.00001, .00002, .00005, .0001, ..., .2, .5, 1]. For the BLSTM-RNN regression models, we adopted two hidden layers. In our experiments, the BLSTM-RNN model consists of 20 nodes per hidden layers with a learning rate of 10^{-5} and a momentum of 0.90. Moreover, zero mean Gaussian noise with a standard deviation of 0.2 was added to the input activations in the training phase to improve generalisation. All weights were randomly initialised in the range from -0.1 to 0.1. The parameters of each model were optimised on the validation set with an early stopping strategy for its corresponding task. To implement the SVR and BLSTM-RNN models, we selected LIBLINEAR [24] and CURRENNT [25] toolkits, respectively, for the sake of reproducibility.

For the baseline, we performed the emotion evaluation on the SVR model and the BLSTM-RNN model independently. Further, to better compare performance between the prediction-based RNN-RNN framework and the traditional deeper RNN framework, we further provided another baseline BLSTM-RNN model with four hidden layers. Whereas, for the proposed methods, two models were successively combined in order to enhance the capability of prediction as demonstrated in Fig. 1. Moreover, as described in Section 3.3, the $Model_2$ in the proposed framework can be trained either with true or pseudo predictions. To simulate pseudo predictions, white Gaussian noise was added to the gold standard with different signal-to-noise ratios, i.e., 0, 3, 6, 9, and 12 dB.

It should be noted that, all feature sets in the training set were standardised to zero mean and unit variance before training $Model_2$, except the initial true predictions or the pseudo predictions. We carried out an on-line standardisation over the validation and the test set, using the means and variances from the corresponding training sets.

Additionally, annotation delay compensation was also performed to compensate for the temporal delay between the observable cues, as shown by the participants, and the corresponding emotion reported by the annotators [26]. As in [16] we identified this delay to be four seconds which was duly compensated, by shifting the gold standard back in time with respect to the features for both arousal and valence in all of our experiments.

To evaluate the agreement level between the predictions of emotion and the gold standard, the standard metric is the Concordance Correlation Coefficient (CCC) [11], which is calculated by

$$\rho_c = \frac{2\sigma_{xy}^2}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{1}$$

Table 2. Concordance Correlation Coefficient (CCC) of the three different hierarchical structures (i.e., RNN-SVR, SVR-RNN, or RNN-RNN) on the validation (val.) and test sets for both arousal and valence regression. For training the second model, either the true *predictions* from the first model or the simulated *pseudo predictions* were used. Baselines and other state-of-the-art results on RECOLA are also presented. The best achieved CCCs are highlighted. The symbol of $*$ indicates the significance of the performance improvement.

CCC	arousal		valence	
	val.	test	val.	test
Baseline				
SVR	0.744	0.726	0.403	0.300
RNN (2 layers)	0.752	0.738	0.329	0.278
RNN (4 layers)	0.747	0.708	0.401	0.305
Models trained with true predictions				
RNN-SVR	0.744	0.726	0.422*	0.387*
SVR-RNN	0.769*	0.730	**0.440***	**0.393***
RNN-RNN	0.757*	0.726	0.418*	0.369*
Models trained with pseudo predictions				
RNN-SVR	0.746	0.729	0.407	0.301
SVR-RNN	**0.774***	0.743*	0.420*	0.373*
RNN-RNN	**0.774***	**0.744***	0.412*	0.377*
State-of-the-art				
CCC-objected [3]	0.412	0.350	0.242	0.199
End-to-End [22]	0.741	0.686	0.325	0.261

with $\mu_x = \mathrm{E}(\mathbf{x})$, $\mu_y = \mathrm{E}(\mathbf{y})$, $\sigma_x^2 = \mathrm{var}(\mathbf{x})$, $\sigma_y^2 = \mathrm{var}(\mathbf{y})$ and $\sigma_{xy}^2 = \mathrm{cov}(\mathbf{xy})$, where $\mathbf{x}$ and $\mathbf{y}$ are two series. In contrast to the largely used Pearson's correlation coefficient, CCC takes also the bias, i.e., $(\mu_x - \mu_y)^2$, into account. Hence, the value of CCC is in range of $[-1, 1]$, where ± 1 represents perfect concordance and discordance while 0 means no correlation.

To further access the significance level of performance improvement, a statistical evaluation was carried out over the whole predictions between the proposed and the baseline methods by means of Fisher's r-to-z transformation [27]. Unless stated otherwise, a p value less than .05 indicates significance.

4.3. Results and Discussion

Table 2 demonstrates the performance of our proposed methods by applying three different hierarchical structures (i.e., RNN-SVR, SVR-RNN, and RNN-RNN) and two training schemes as mentioned in Section 3.3, evaluating on the validation and the test sets for both the arousal and valence dimensions in terms of CCC. The parameters of the complexity value of SVR and the SNR of the simulated predictions were optimised on the validation set, which were then applied for the test set.

Generally speaking, the proposed methods outperform the corresponding baselines in most cases on the validation set for both arousal and valence. Regarding the test set, the proposed methods performed better than the baseline in some cases for arousal and in all cases for valence. Moreover, we could observe that the improvement on valence is much more than that on arousal. This may be due to the fact that, the models have already fit well for the arousal with the original feature vector.

Furthermore, comparing RNN-SVR and SVR-RNN, one may observe that, the performance of the latter is on average higher than the former. This is possibly because SVR could help the following RNN to avoid the local minima by using its prediction. Whereas, the local minimum problem of the RNN in the RNN-SVR framework could not be eased by the SVR model followed it. This might imply that the order of the models in the hierarchical structure matters and the order should be chosen with care. Additionally, when comparing the two different training strategies, the pseudo predictions performs competitive to or even slightly better than the true predictions, whilst it does not require any prior knowledge of the first model.

Comparing the cascaded RNN-RNN structure and the deeper RNN with four layers, one may observe that, simply adding more hidden layers to make the RNN deeper, does not lead to the same improvement as the cascaded RNN-RNN in which the output predictions of the first shallow RNN model jointed with other features pairwise are fed into a second shallow RNN model. A rationale behind this is that the gradient of the neural network is supposed to be vanished when simply increasing the hidden layers. The error information can not efficiently back-propagate to the low layers near the visual input layer. Meanwhile, the data information at the low layers can also not be accessed well by the high layers that are far from the visual layer [18]. By our method, however, the data information can be explicitly transferred to the hidden layers in the later RNNs by means of the predictions from the former RNNs.

Particularly, the best performance on the test set is obtained at .744 of CCC for arousal when using RNN-RNN trained with pseudo predictions, and .393 of CCC for valence by SVR-RNN trained with the original SVR predictions. These results are also much better than the ones achieved by using a novel discriminative training approach based on CCC [3], and the ones by using novel features extracted from the raw waveform by performing convolutional neural networks [22], respectively.

5. CONCLUSION

Various regression models have been successfully applied independently or associatively for different tasks, particularly in continuous affective computing. In this paper, we presented a prediction-based learning framework which associates different regression models to leverage the performance of continuous prediction of emotion. The proposed method takes advantages of multiple models, such as the context-sensitive capability of memory-enhanced neural networks, and the global-optimisation capability of Support Vector Regression. By implementing the experiments on a spontaneous database, significant improvement of performance indicates the effectiveness of our methods.

Although the proposed framework aims to exploit the advantages of various models, it is sometimes inevitable to take the disadvantages, which may results in a negative effect to the framework. In future, more efforts will be paid to address this problem. Further, the proposed framework will be evaluated by other applications (e. g., text and video) and other regression models.

6. ACKNOWLEDGEMENTS

This work was partially supported by the EC's 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), the EU's Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA) and the Research Innovative Action No. 645378 (ARIA-VALUSPA), and by the German Federal Ministry of Education, Science, Research and Technology (BMBF) under grant agreement No. 16SV7213 (EmotAsS).

7. REFERENCES

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

[2] S. Petridis and M. Pantic, "Prediction-based audiovisual fusion for classification of non-linguistic vocalisations," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 45–58, Jan 2016.

[3] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, New York, NY, 2016, pp. 2196–2202.

[4] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," *Advances in Neural Information Processing Systems*, vol. 9, pp. 155–161, 1997.

[5] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, Sep 2001.

[6] T.-Y. Kwok and D.-Y. Yeung, "Constructive algorithms for structure learning in feedforward neural networks for regression problems," *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp. 630–645, May 1997.

[7] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jul 2006.

[8] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, Mar 1989.

[9] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, Apr 2011.

[10] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *Proc. Affective Computing and Intelligent Interaction (ACII)*, Xi'an, China, 2015, pp. 698–704.

[11] 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. International Workshop on Audio/Visual Emotion Challenge*, Brisbane, Australia, 2015, pp. 3–8.

[12] J. Wei, E. Pei, D. Jiang, H. Sahli, L. Xie, and Z. Fu, "Multimodal continuous affect recognition based on LSTM and multiple kernel learning," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Siem Reap, Cambodia, 2014, pp. 1–4.

[13] X. Qiu, L. Zhang, Y. Ren, P. N. Suganthan, and G. Amaratunga, "Ensemble deep learning for regression and time series forecasting," in *Proc. Computational Intelligence in Ensemble Learning (CIEL)*, Orlando, FL, 2014, pp. 1–6.

[14] M. Wöllmer, F. Weninger, J. Geiger, B. Schuller, and G. Rigoll, "Noise robust ASR in reverberated multisource environments applying convolutive NMF and Long Short-Term Memory," *Computer Speech & Language*, vol. 27, no. 3, pp. 780–797, May 2013.

[15] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, Mar 2012.

[16] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proc. the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Brisbane, Australia, 2015, pp. 41–48.

[17] S. R. Gunn, "Support vector machines for classification and regression," School of Electronics and Computer Science, University of Southampton, Southampton, England, Tech. Rep. 14, May 1998.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov 1997.

[19] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, Jul 2005.

[20] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. EmoSPACE (FG)*, Shanghai, China, 2013, pp. 1–8.

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

[22] G. Trigeorgis, F. Ringeval, R. Bruckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a Deep Convolutional Recurrent Network," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5200–5204.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM International Conference on Multimedia (ACM MM)*, Florence, Italy, 2010, pp. 1459–1462.

[24] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, Jun 2008.

[25] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT: The munich open-source cuda recurrent neural network toolkit," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 547–551, Jan 2015.

[26] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, Apr 2015.

[27] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*. Abingdon, UK: Routledge, 2013.