END-TO-END LEARNING FOR DIMENSIONAL EMOTION RECOGNITION FROM PHYSIOLOGICAL SIGNALS

Gil Keren[1], Tobias Kirschstein[1], Erik Marchi[1,2*], Fabien Ringeval[1,3], Björn Schuller[1,4]

[1] Chair of Complex & Intelligent Systems, University of Passau, Germany
[2] Apple Inc.
[3] Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, France
[4] Department of Computing, Imperial College London, UK.

ABSTRACT

Dimensional emotion recognition from physiological signals is a highly challenging task. Common methods rely on hand-crafted features that do not yet provide the performance necessary for real-life application. In this work, we exploit a series of convolutional and recurrent neural networks to predict affect from physiological signals, such as electrocardiogram and electrodermal activity, directly from the raw time representation. The motivation behind this so-called *end-to-end* approach is that, ultimately, the network learns an intermediate representation of the physiological signals that better suits the task at hand. Experimental evaluations show that, this very first study on *end-to-end* learning of emotion based on physiology, yields significantly better performance in comparison to existing work on the challenging RECOLA database, which includes fully spontaneous affective behaviors displayed during naturalistic interactions. Furthermore, we gain better understanding of the models' inner representations, by demonstrating that some cells' activations in the convolutional network are correlated to a large extent with hand-crafted features.

Index Terms— End-to-end learning, Physiological signals, Emotion recognition, Convolutional Neural Networks, Long Short-Term Memory Recurrent Neural Networks

1. INTRODUCTION

The automatic recognition of affective behaviors has received a significant increase of attention in the last decade, both from industry and academics. Indeed, emotion plays a major role in many key aspects of everyday life interactions, such as rational decision-making, collaborative work, learning, and health care. Physiological signals are supposed to provide relevant insights on emotion, as they are correlated with responses of the autonomic nervous system, which can be produced during adaptation to the environment or to emotional stimuli [1]. However, such signals are not directly perceptible the way audiovisual are. While emotion recognition from vocal and facial expressions has matured enough in the last decade to approach real-life applications [2], performance achieved on peripheral physiological signals, such as electrocardiogram (ECG) and electrodermal activity (EDA), has not yet lead to satisfactory results. Those two signals can nevertheless provide complementary descriptions of spontaneous emotion in multimodal fusion frameworks [3, 4].

One of the advantages of emotion monitoring from physiological signals, in comparison with audiovisual, is that their acquisition is done unconsciously; they also require much less energy and storage capacity, thus allowing emotion sensing 'in the wild' at a reduced cost. However, measurements of those signals are prone to errors due to movements, and are also subject to non-stationary variations that are independent of emotion. Therefore, there is a critical need for research on how peripheral physiological signals can be exploited to perform robust prediction of spontaneous affective behaviours.

Whereas systems from the literature have relied on hand-crafted features to perform emotion sensing so far, we propose in this paper a radically different approach: the raw signals are directly fed into deep neural networks that perform the so-called *end-to-end* learning of the emotion. The motivation behind this idea is that, ultimately, the network learns an intermediate representation of the raw input that better suits the task at hand, and hence leads to improved performance [5, 6]. The main contributions of this paper are the following: we perform the first attempt in *end-to-end* learning of peripheral physiological signals and apply it to dimensional emotion recognition. We also show that, this method can yield a large improvement in performance over hand-crafted features on the challenging RECOLA database [7], and that a small proportion of the representations learned by the network presents a high correlation with hand-crafted features.

The remainder of this paper is structured as follows: related work on emotion recognition from ECG and EDA, and existing attempts on *end-to-end* learning from audiovisual signals are described in section 2, details of our method are explained in section 3, evaluations are carried out in section 4, and a conclusion is given in section 5.

2. RELATED WORK

There exists many peripheral physiological signals that can be exploited to sense human emotions, e. g., respiration amplitude [8] and pupillary response [9]. In this paper, we focus on the ECG and EDA signals, as they can nowadays easily be captured with wearable devices, such as smart-watches and smart-bracelets, thus allowing emotion monitoring 'in the wild' [10, 11] and continuously over the day if desired or needed. Moreover, it is worth to mention that non-contact methods can also be used to estimate both ECG and EDA signals from video data [12], or even from audio [13] or motion sensors included in smartphones [14].

The last two editions of the Audio Visual Emotion recognition Challenge (AVEC) [3, 4], included a task on dimensional emotion recognition from audiovisual and physiological signals, using the same database of this study. We therefore describe in the following

*Now with Apple Inc. The work was done while at [1] and it does not contain Apple proprietary information.

section the best performing systems developed by the participants of the AVEC challenges, cf. Table 1. The evaluation metric used for the (speaker-independent) dimensional emotion prediction task is the concordance correlation coefficient (CCC) [15], computed over concatenated instances [16]:

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

where ρ is the Pearson's correlation coefficient between the predictions and the emotion labels, μ_x, μ_y are the means of the two signals and σ_x, σ_y their standard deviations. CCC thus combines the Pearsons correlation coefficient with the square difference between the mean of the two compared time series, which makes it sensitive to bias and scaling factors [17].

The overview of the three best performing systems of the last two editions of the AVEC challenge shows that, hand-crafted features computed from the ECG clearly outperform those extracted from the EDA for both arousal and valence. The best performance achieved on arousal has been obtained with the system developed by Weber et al. [21], using the baseline feature set of the challenge; 19 features composed of linear and non-linear descriptors computed on a band-pass filtered version of the ECG signal. A Support Vector Regression (SVR) model was trained for each subject of the database, and fusion of these single-speaker-regression-models was performed by a linear regression. This approach has also provided the best performance on arousal for all other signals excepted EDA, and on valence with the heart rate descriptors (HRHRV); five statistical measures computed on the R-R signal and its first order derivative. The interest of combining single-speaker-regression-models to adapt the system to the speaker peculiarities has also been recently demonstrated for acoustic features [24]. In fact, methods based on *end-to-end* learning have been applied on the RECOLA database, using either the raw acoustic waveform [6] or the video signal [5], yet not physiological information as pursued here. Performance reported with this approach shows that the system can learn intermediate representations of the data that are related to affective behaviours, and even outperform methods applied on hand-crafted features, thus demonstrating the importance of this approach.

3. METHOD

An overview of the main building blocks of the proposed system's architecture is given in Figure 1. It incorporates preprocessing phases for downsampling, normalisation and windowing of the input signal. The short-term views of the signals are fed into zero or more convolutional blocks, each convolutional block being comprised of a convolutional layer with a kernel size greater than one, a non-linear activation function, a convolution layer with kernel size of one, another non-linear activation function, and a max-pooling layer applied over windows of two time-steps. The output of the convolutional blocks is fed into zero or more recurrent layers, and their output into one or more fully connected layers with a non-linear activation function, followed by the output layers described in Section 3.7. An appropriate loss function is computed from the output layers, that is in turn minimised by a gradient-based optimisation algorithm, cf. Section 4 for specifications. Some components of the model are optional, such as the regularisation techniques described in Section 3.6 and batch normalisation.

3.1. Normalisation and downsampling

The input signal of the model is a real valued signal x, which consists of k channels at each time-step. We normalise the mean and standard deviation of each channel separately, across all time-steps. The signal is downsampled to a target-frequency f_m in order to remove redundant samples. We match the frequency f_y of the gold-standard (computed from a pool of time-continuous annotations) in the training set to f_m, by downsampling in case $f_y > f_m$, or by applying linear extrapolation in case $f_y < f_m$. As in some cases the input signal might encode variations in the annotation data better than it encodes the actual values, we center the gold-standard during the training phase, and restore those mean values during the inference phase.

3.2. Windows extraction

Once normalisation and downsampling is done, the frequency of the input and output signals is matched. We extract a maximal number of overlapping windows of s seconds from each input signal in the training set, each is paired with the annotation that corresponds to the center of the window; input signals are padded on the edges. The extracted window-annotation pairs from the training set are the training examples the model is trained on. Note that, by extracting a maximal number of windows (windows are shifted by one time-step from each other), we augment our training set with the aim to improve the performance of the model [25]. For the evaluation set, windows are extracted with the same procedure as for the training set, and we downsample/interpolate the time-sequence of network predictions for the different windows, to match the frequency f_y of the output signal.

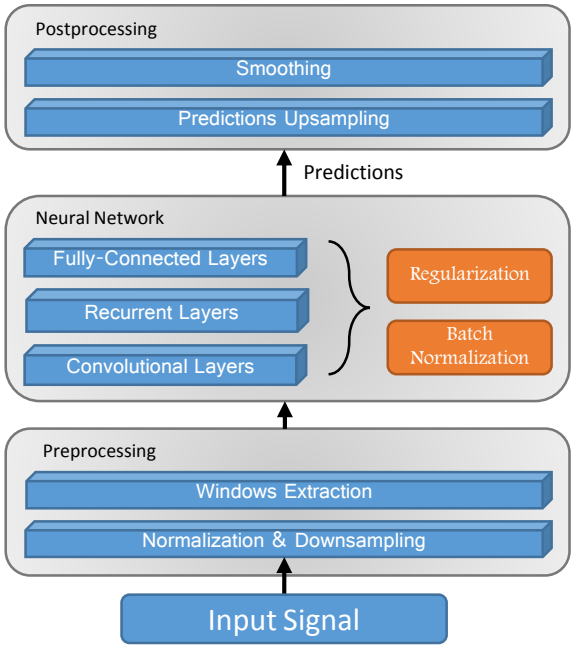

Fig. 1. Main building blocks of the proposed deep end-to-end learning method.

Table 1. Comparison of performance (CCC) for speaker independent dimensional emotion recognition from various peripheral physiological signals according to the best three systems of the last two editions of the AVEC challenge (development partition). Note that for AVEC'15, HRHRV features were included in the ECG feature set, and SCL and SCR features were included in the EDA feature set. Best performance obtained on each channel is highlighted in bold style for arousal and valence; HRHRV: heart rate and heart rate variability; SCL: skin conductance level; SCR: skin conductance resistance.

	Arousal					*Valence*				
Authors	ECG	HRHRV	EDA	SCL	SCR	ECG	HRHRV	EDA	SCL	SCR
AVEC'15 [3]	.275	—	.078	—	—	.183	—	.204	—	—
Kächele et al. [18]	.344	—	.125	—	—	.256	—	.236	—	—
Chen et al. [19]	.333	—	—	—	—	**.314**	—	.315	—	—
He et al. [20]	.297	—	**.248**	—	—	.293	—	.231	—	—
AVEC'16 [4]	.271	.379	.073	.068	.073	.153	.293	.194	.166	.085
Weber et al. [21]	**.468**	**.424**	.187	**.197**	**.193**	.221	**.413**	.281	.277	.174
Povolný et al [22]	.323	.391	.123	.134	.167	.272	.388	**.316**	**.310**	**.194**
Sun et al. [23]	.320	.392	.122	.116	.117	.167	.264	.234	.229	.126

3.3. 1D convolutional layers

A 1D convolutional layer [26] processes a time-sequence with k channels, by convolving windows of l time-steps with learnable kernels of size $l \times k$ each. When a convolutional layer is fed with a time-sequence, it performs a convolution of the time-sequence with n different such kernels resulting in a time-sequence of the same length as the original time-sequence and n channels. A fixed channel across all time-steps is called a *feature map*. Whereas a 1D max-pooling layer processes each channel separately by performing a *max* operation across all time-steps of a defined window. When max-pooling layers is applied on a time-sequence using a window size of l and a stride of d, each time-step in the resulting time-sequence corresponds to a window in the original time-sequence, where the windows are shifted with d time-steps between them.

3.4. Recurrent layers

Long-term dependencies over time are modelled with recurrent layers. Given a time-sequence $a = (a_1, \ldots, a_t)$, a recurrent layer generates a sequence of hidden states $h = (h_1, \ldots, h_t)$ by performing $h_{t+1} = RNN(h_t W_h + a_{t+1} W_a + b)$, where W_h, W_a are learnable weight matrices, b is a learnable bias, and RNN is a transition function that depends on the type of recurrent layer used. The last m elements of the time-sequence h are then passed to the next layer. One natural extension of a recurrent layer is a bidirectional recurrent layer, that consists of two recurrent layers as defined above. In this approach, the first recurrent layer processes the input time-sequence a forward from a_1 to a_t, while the other processes it backwards, from a_t to a_1.

3.5. Batch normalisation

When training neural networks, the distribution of each layer's inputs changes during training, as the parameters of the previous layers change. This phenomenon slows down training by requiring smaller learning rates and careful parameter initialisation. To alleviate this issue, batch normalisation [27] has been proposed. It consists in applying a linear transformation on the output of a layer – just before applying the activation function – that enforces learnt values of the mean and standard deviation of each unit / feature map in the output. The mean and standard deviation are calculated per mini-batch of examples.

3.6. Regularisation

In order to reduce overfitting, which is a well known issue when training neural networks on relatively small sized datasets, different regularisation techniques can be applied. A prominent regularisation method for neural networks is dropout [28], that when applied on some layer of a neural network, each element in the – possibly multidimensional – output is set to zero with probability p or else multiplied by $\frac{1}{1-p}$. At inference time, dropout is not used. Other regularisation techniques used in our models are L1/L2 weight decay and adding a zero-mean Gaussian noise to the network's input.

3.7. Regression through classification

For the prediction of continuous-valued emotion labels from windows of the input signal, the last layer of the network is usually a standard fully-connected layer with one output unit per emotion dimension; multi-task learning can be performed here by using more than one output unit [16]. Alternatively, the continuously-valued emotion labels can be discretised into a number of classes, where a softmax layer is then employed to model the output distribution. In this study, we use this regression through classification approach, with a linear discretisation of the labels to fit the range of continuous values into the desired number of classes.

4. EXPERIMENTS

The database used in this study is briefly described in the following section. We then discuss the optimisation of the hyper-parameters and architecture of the system. Subsequently, obtained results are described followed by an analysis of the representations learnt by the *end-to-end* system.

4.1. RECOLA database

The Remote Collaborative and Affective Interactions (RECOLA) database [7] contains spontaneous and naturalistic dyadic interactions of French-speaking adults during the resolution of a collaborative task. Multi-modal signals, i. e., audio, video, ECG and EDA, were continuously and synchronously recorded from 27 French-speaking subjects. Time-continuous ratings (40 ms binned frames) of emotional arousal and valence were created for the first five minutes of all recordings. Those ratings were averaged over six raters to create a single time-continuous emotion label for each dimension

Table 2. Chosen target-frequencies f_m and windows sizes s.

	Signal	f_m [Hz]	s [seconds]
Arousal	EDA	1	75
	ECG	25	6
	HR	25	8
	SCL	1.25	12
	SCR	2.5	12
Valence	EDA	1	50
	ECG	25	8
	HR	25	14
	SCL	5	16
	SCR	5	15

[4]. For experimental evaluations, the dataset is equally divided into speaker-disjoint subsets for training, development (validation) and testing.

For the purpose of the AVEC'16 challenge [4], the heart-rate (HR) derived from the ECG was provided as an additional physiological signal. Regarding EDA, the skin conductance response (SCR) and skin conductance level (SCL) signals were also extracted and provided as separate physiological descriptors. Note that for EDA, SCR and SCL, test data from subject #7 was not used, due to an issue during the recording of this subject.

4.2. Hyperparameters and architectural choices

We optimised over many architectural setups and hyperparameters using the training and development sets, for every combination of input-signal (i. e., ECG, HR, EDA, SCL, and SCR) and emotional dimension. The best performing models were then used for computing predictions on the test set. Regarding the preprocessing methods, we experimented with different granularity with the target-frequency f_m and window size s, cf. Sections 3.1 and 3.2, respectively, as well as with shifting the mean of the gold-standard to zero. The values of the best performing target-frequency and window size for each physiological signal and emotional dimension is given in Table 2. Results show that, ECG and HR signals perform best when they are processed with the same time-granularity as the one used on the gold-standard (25 Hz), and that their best window size is in the same ballpark as those exploited for audiovisual data [6], whereas all three EDA related signals perform best with a low target-frequency and much longer window size, especially for EDA.

For the model architecture, we optimised the number of convolutional, recurrent and fully-connected layers, and the number of units / feature maps in each layer, as well as the size of the convolution kernel, type of recurrent layers (LSTM [29] / BLSTM [30]), and number of last time-steps from the recurrent layers to use as the layer's output. In addition, we performed optimisation on the regularisation strategy, by experimenting with different levels of L1/L2 weight decay, Gaussian noise on the input, layers to apply dropout on and dropout probability. The choice of learning algorithm was also optimised (Stochastic Gradient Descent and Adadelta [31]), with different learning rates and a mini-batch size of 100 window-annotation pairs. Regarding discretisation of the labels to perform regression through classification task, we used 61 classes, with values between -0.30 and 0.30, and 0.01 shifts. Cross-entropy loss was used as loss function in the training of models with discritised output, whereas a CCC loss function was used in the other models [6, 17, 18]. The type of non-linearity in the convolutional and fully-connected layers, and batch normalisation were also optimised.

Finally, for all investigated methods, a chain of post-processing is applied to the predictions obtained on the validation set, as done in the AVEC baseline system [4]: (i) median filtering, with size of the window ranging from 0.4 s to 20 s, (ii) centring by computing the bias between gold-standard and prediction, and (iii) scaling, by using the ratio of standard-deviation of gold-standard and prediction as scaling factor. Any of these post-processing steps is kept when an improvement is observed on the CCC of the validation set, and applied then with the same configuration on the test partition.

Table 3. Comparison of performance (CCC) for arousal recognition from the various physiological signals. The proposed method (End-to-End), the AVEC2016 baseline [4] and the two best submissions to the AVEC2016 challenge [21, 22].

	Signal	End-to-End	[4]	[21]	[22]
Dev	ECG	.267	.271	**.468**	.323
	HR	**.426**	.379	.424	.391
	EDA	**.212**	.073	.187	.123
	SCL	.149	.068	**.197**	.134
	SCR	.189	.073	**.193**	.167
Test	ECG	**.309**	.158	—	—
	HR	**.360**	.334	—	—
	EDA	**.101**	.075	—	—
	SCL	**.190**	.066	—	—
	SCR	**.257**	.065	—	—

Table 4. Comparison of performance (CCC) for valence recognition from the various physiological signals. The proposed method (End-to-End), the AVEC2016 baseline [4] and the two best submissions to the AVEC2016 challenge [21, 22].

	Signal	End-to-End	[4]	[21]	[22]
Dev	ECG	.135	.153	.221	**.272**
	HR	**.419**	.293	.413	.388
	EDA	.284	.194	.281	**.316**
	SCL	.308	.166	.277	**.310**
	SCR	**.286**	.085	.174	.194
Test	ECG	**.210**	.121	—	—
	HR	**.225**	.198	—	—
	EDA	**.336**	.228	—	—
	SCL	**.353**	.216	—	—
	SCR	**.313**	.145	—	—

4.3. Quantitative Results

We evaluated our deep neural network models for the prediction of the two AVEC challenge dimensions of emotion from the RECOLA dataset: arousal and valence. Our models' performance is compared to the best results from the submissions to the AVEC challenge and to the challenge baseline itself. Results for predicting arousal and valence levels on the development and test sets from each of the different input-signals used are presented in Tables 3 and 4. The results demonstrate that in terms of concordance correlation coefficient (CCC), the official evaluation metric for this dataset, our end-to-end models in almost all settings outperformed the AVEC2016 baseline, and yielded superior or comparable performance to the best submissions thereof, that use traditional hand-crafted features with a much

Table 5. Multi-modal predictions (in CCC)

	Arousal	Valence
Dev	.463	.477
Test	.430	.407

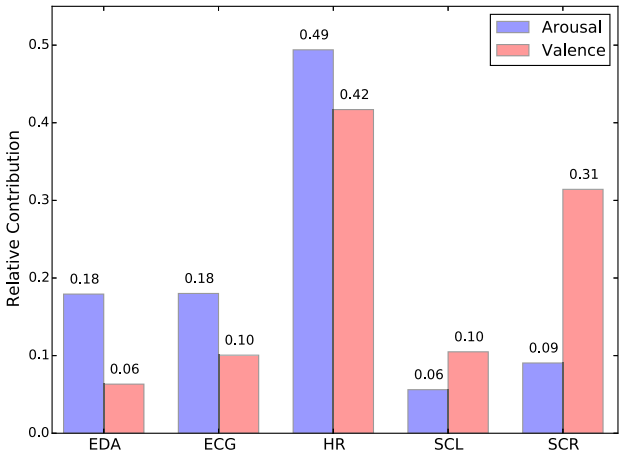

Fig. 2. Contributions of the different signals to the final prediction in the late fusion model.

more complicated model pipeline. The results therefore support our initial hypothesis, that an end-to-end learning approach directly from raw signals can be beneficial for this task.

Multimodal fusion of the five modalities (EDA, ECG, HR, SCL, SCR) was performed using the same procedure as the procedure as in the AVEC2016 baseline. We employed a late-fusion scheme with linear regression:

$$Pred_{muilti} = \beta_0 + \sum_{i=1}^{5} \beta_i Pred_i,$$

where $Pred_i$ is the prediction-signal using modality i, and $\{\beta_0, \ldots, \beta_5\}$ are real-valued scalars that are optimised using the development set predictions. As expected, the results in Table 5 confirm that fusing the predictions from the different signals further improves the CCC measure. The learned coefficients $\beta_1, \ldots \beta_5$ allow us to infer about the relative importance of each independent signal for the prediction of each emotion dimension. In order to depict the contribution of each modality in the prediction, we normalised the learned linear regression coefficients that were learnt for the multimodal fusion model into percentage: $C_i = |\beta_i| / \sum_{i=1}^{5} \beta_i$. The normalised coefficients $C_1, ..., C_5$ are then depicted in Figure 2. It presents that the HR signal was the most dominant input-signal for the prediction of both arousal and valence. The SCR signal contributed more to the prediction of valence, while the EDA signal contributed mostly to the arousal prediction.

4.4. Relation to Hand-Crafted Features

We investigated different cell activations (outputs) in the convolutional layers of the best performing models to gain a better understanding of the internal representations that our models learnt. We found cells with activations that highly correlate with some hand-crafted features that were extracted for the AVEC2016 challenge and that were used by the baseline approach to predict arousal and valence. Some examples can be seen in Figure 3. This demonstrates the feature extraction capabilities of convolutional neural networks, as these models are able to learn relevant and interpretable features solely from data, when trained in an end-to-end manner.

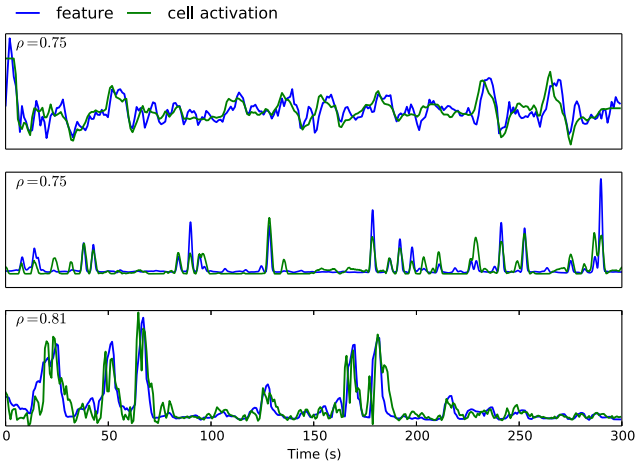

Fig. 3. Examples of correlations between cell activations in our neural network models and hand-crafted features from the AVEC2016 baseline (mean and standard deviation are normalized). From top to bottom: EDA skewness, ECG spectal coefficient #3, SCL standard deviation derivative.

5. CONCLUSION

In a first of its kind study, we successfully employed a series of convolutional and recurrent neural networks to predict levels of arousal and valence from physiological signals, such as electrocardiogram and electrodermal activity, directly from the raw time representation. Experimental evaluation shows that in almost all settings, our end-to-end approach yields superior performance to the AVEC2016 baseline and superior or comparable results to the strongest baseline systems, that use hand-crafted features and employ a fairly complicated model pipeline. Furthermore, we show that some cells' activations in the network are correlated to a large extent with hand-crafted features, thus gaining better understanding of our models' inner representations. In future work, we plan to continue exploring the advances of deep neural networks paired with raw features, to further improve the prediction ability of dimensional emotion.

6. ACKNOWLEDGMENTS

This work has been supported by the European Union's Seventh Framework Programme through the ERC Starting Grant No. 338164 (ERC StG iHEARu).

7. REFERENCES

[1] R. Lazarus, *Emotion and adaptation*, Oxford University Press, New York, 1991.

[2] F. Eyben, M. Unfried, G. Hagerer, and B. Schuller, "Automatic Multi-lingual Arousal Detection from Voice Applied to Real Product Testing Applications," in *Proc. of ICASSP*, New Orleans, LA, 2017.

[3] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data," in *Proc. of AVEC, ACM MM*, Brisbane, Australia, 2015, pp. 3–8.

[4] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proc. of AVEC'16, ACM MM*, Amsterdam, The Netherlands, 2016, pp. 3–10.

[5] Pooya Khorrami, Tom Le Paine, Kevin Brady, Charlie Dagli, and Thomas S Huang, "How deep neural networks can improve emotion recognition on video data," in *Proc. of ICIP*, Phoenix, AZ, 2016, pp. 619–623.

[6] G. Trigeorgis, F. Ringeval, R. Brueckner, 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. of ICASSP*, Shanghai, China, 2016, pp. 5200–5204.

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

[8] F. A. Boiten, N. H. Frijda, and C. J. E. Wientjes, "Emotions and respiratory patterns: review and critical analysis," *International Journal of Psychophysiology*, vol. 17, no. 2, pp. 103–128, 1994.

[9] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 511–677, 2008.

[10] V. Alexandratos, M. Bulut, and R. Jasinschi, "Mobile real-time arousal detection," in *Proc. of ICASSP*, Florence, Italy, 2014, pp. 4427–4431.

[11] A. Bachmann, C. Klebsattel, A. Schankin, T. Riedel, M. Beigl, M. Reichert, P. Santangelo, and U. Ebner-Priemer, "Leveraging smartwatches for unobtrusive mobile ambulatory mood assessment," in *Proc. of UbiComp/ISWC*, Osaka, Japan, 2015, pp. 1057–1062.

[12] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[13] B. Schuller, F. Friedmann, and F. Eyben, "Automatic Recognition of Physiological Parameters in the Human Voice: Heart Rate and Skin Conductance," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7219–7223.

[14] J. Hernandez, D. J. McDuff, and R. W. Picard, "Biophone: Physiology monitoring from peripheral smartphone motions," in *Proc. of EMBC*, Milano, Italy, 2015, pp. 7180–7183.

[15] I-Kuei L. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[16] 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, 2015.

[17] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller, "Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio," in *Proc. of IJCAI*, New York City, NY, 2016, pp. 2196–2202.

[18] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels, "Ensemble methods for continuous affect recognition: multimodality, temporality, and challenges," in *Proc. of AVEC, ACM MM*, Brisbane, Australia, 2015, pp. 9–16.

[19] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proc. of AVEC, ACM MM*, Brisbane, Australia, 2015, pp. 49–56.

[20] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. of AVEC, ACM MM*, Brisbane, Australia, 2015, pp. 73–80.

[21] R. Weber, V. Barrielle, C. Soladié, and R. Séguier, "High-level geometry-based features of video modality for emotion prediction," in *Proc. of AVEC, ACM MM*, Amsterdam, The Netherlands, 2016, pp. 51–58.

[22] F. Povolný, P. Matějka, M. Hradiš, A. Popková, L. Otrusina, and P. Smrž, "Multimodal emotion recognition for AVEC 2016 challenge," in *Proc. of AVEC, ACM MM*, Amsterdam, The Netherlands, 2016, pp. 75–81.

[23] B. Sun, S. Cao, L. Li, J. He, and L. Yu, "Exploring multimodal visual features for continuous affect recognition," in *Proc. of AVEC, ACM MM*, Amsterdam, The Netherlands, 2016, pp. 83–88.

[24] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. Di Natale, "Continuous estimation of emotions in speech by dynamic cooperative speaker models," *IEEE Transactions on Affective Computing*, 2016, in press.

[25] G. Keren, J. Deng, J. Pohjalainen, and B. Schuller, "Convolutional neural networks with data augmentation for classifying speakers native language," in *Proc. of INTERSPEECH*, San Francisco, CA, 2016, pp. 2393–2397.

[26] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of ICML*, Lille, France, 2015, pp. 448–456.

[28] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

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

[30] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[31] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.