Pattern Recognition Letters
journal homepage: www.elsevier.com

Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data

Fabien Ringeval[a,d,**], Florian Eyben[a], Eleni Kroupi[b], Anil Yuce[c], Jean-Philippe Thiran[c], Touradj Ebrahimi[b], Denis Lalanne[d], Björn Schuller[a,e,f]

[a] *TU München, Machine Intelligence & Signal Processing group, MMK, Arcisstrasse 21, 80333 München, Germany*
[b] *Ecole Polytechnique Fédérale de Lausanne, Multimedia Signal Processing Group, Station 11, 1015 Lausanne, Switzerland*
[c] *Ecole Polytechnique Fédérale de Lausanne, Signal Processing Laboratory (LTS5), Station 11, 1015 Lausanne, Switzerland*
[d] *Université de Fribourg, Document Image and Voice Analysis, DIVA, Boulevard de Pérolles 90, 1700 Fribourg, Switzerland*
[e] *Imperial College London, Department of Computing, Machine Learning Group, 180 Queen's Gate, SW7 2AZ London, U. K.*
[f] *University of Passau, Chair of Complex Systems Engineering, Innstrasse 33, 94032 Passau, Germany*

ABSTRACT

Automatic emotion recognition systems based on supervised machine learning require reliable annotation of affective behaviours to build useful models. Whereas the dimensional approach is getting more and more popular for rating affective behaviours in continuous time domains, e. g., arousal and valence, methodologies to take into account reaction lags of the human raters are still rare. We therefore investigate the relevance of using machine learning algorithms able to integrate contextual information in the modelling, like Long Short-Term Memory Recurrent Neural Networks do, to automatically predict emotion from several (asynchronous) raters in continuous time domains, i. e., arousal and valence. Evaluations are performed on the recently proposed RECOLA multimodal database (27 subjects, 5 minutes of data and 6 raters for each), which includes audio, video, and physiological (ECG, EDA) data. In fact, studies uniting audiovisual and physiological information are still very rare. Features are extracted with various window sizes for each modality and performance for the automatic emotion prediction is compared for both different architectures of Neural Networks and fusion approaches (feature-level / decision-level). The results show that: (i) LSTM network can deal with (asynchronous) dependencies found between continuous ratings of emotion with video data, (ii) the prediction of the emotional valence requires longer analysis window than for arousal and (iii) a decision-level fusion leads to better performance than a feature-level fusion. The best performance (concordance correlation coefficient) for the multimodal emotion prediction is 0.804 for arousal and 0.528 for valence.

1. Introduction

In everyday social interactions, humans express various complex feelings through several communication modalities, such as voice, face, body and even physiology (e. g., sweating). Despite the fact that cognitive processes used to encode affective information during such social interactions are relatively complex, humans can easily manage to decode such information in real time from multimodal cues. However, when such task has to be performed by a machine, e. g., for enabling more natural interactions between human and machines, the complexity of the affect decoding process requires the use of a combination of sophisticated processes of signal processing methods (for extracting relevant information from recordings) and machine learning algorithms (for finding the underlying emotion). Automatic emotion recognition also requires to define models of emotion that can be used for learning the associations between measurable events of various timings (e. g., speech, facial or gestural behaviour) and the corresponding emotion. Human ratings of emotion, performed on collections of audiovisual recordings, can be used to create such dictionary, which can be then used by machines to perform automatic predictions. Two main approaches have been employed so far to per-

[**] Corresponding author: Tel.: +49-89-289-28562; fax: +49-89-289-28535; *e-mail:* fabien.ringeval@tum.de (Fabien Ringeval)

form the ratings of emotion: a categorical approach, where a human rater selects an emotionally related adjective from a list of adjectives to describe the stimuli, or a dimensional approach, where the rater continuously evaluates the emotion observed in the stimuli, using predefined scales, such as arousal and valence. However, dimensional annotations of emotion pose several challenges for being used by machines for automatic emotion predictions (Gunes and Schuller (2013)). While multiple raters are required to increase the reliability of the annotated emotion, the complexity of defining a gold standard from the pool of raters also increases. Indeed, humans have natural bias and inconsistencies in their judgement (Tversky (1969)), which creates some noise in the ratings. They also need time to feedback on the perceived cues, which leads to a lag in time between the observable event and the reported emotion. Whereas noise in the ratings can be filtered out by averaging over (many) different raters, the lag in emotion feedback has important consequences on machine learning algorithms that use fixed window analysis. Indeed, measurable events are shifted in time with the corresponding emotion. Therefore, compensation techniques based on signal processing were investigated in order to define a reliable gold standard of emotion from time-continuous annotations (Nicolaou et al. (2010b); Nicolle et al. (2012); Mariooryad and Busso (2013, 2014)). Such methods require, however, a first analysis of the ratings along with the recordings, to estimate the reaction time of the raters, before using traditional machine learning techniques, such as Support-Vector Machines (SVM), to perform automatic predictions. Yet, the recent apparition of machine learning algorithms able to integrate contextual information could enable the possibility to directly use the information provided by several raters in the system, without using techniques compensating for different reaction time-delays between the raters.

The major contributions investigated in this study can be listed as follow: (i) we investigate the relevance of machine learning algorithms able to integrate contextual information, such as Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) do, to perform automatic prediction of emotion on ratings provided by several raters, (ii) we perform this evaluation on a fully naturalistic multimodal database (RECOLA) that contains audiovisual alongside physiological affective data, (iii) we study the influence of various window sizes to predict emotion from different modalities (i. e., audio, video, ECG and EDA), (iv) we perform multimodal fusion of these modalities at two levels: features or decision and (v) evaluate the interest of using event based probability (e. g., speech utterance or visible face) as a feature to learn which modality to trust over time.

In the remainder of this paper we introduce related work on automatic emotion recognition (Sec. 2), then present the methods used in this study (Sec. 3), and the results (Sec. 4) before concluding (Sec. 5).

2. Related work

In affect recognition, most methods that include speech as a modality have dealt with recognition on an ‘utterance’ level, e. g., Ververidis and Kotropoulos (2006); Vlasenko et al. (2007). Thereby each utterance has exactly one affective label and the classifier or regressor returns exactly one prediction for the utterance. Recently, databases with time-continuous ratings have emerged such as the Sensitive Artificial Listener (SAL) set in the HUMAINE database (Douglas-Cowie et al. (2007)), and the SEMAINE database (Schröder et al. (2012)). Such databases have caused a shift in methods, first of all moving from classification to regression to be able to model continuous affective dimensions (Grimm et al. (2007)), and next moving from utterance or segment level labels (Chetouani et al. (2009)) to quasi time-continuous labels (Eyben et al. (2010a); Schröder et al. (2012)). Automatic emotion recognition from time-continuous labels presents however several challenges that are summarised below.

It requires the determination of the appropriate length of the temporal window used for emotion prediction, which depends on the modality and the emotion (Gunes and Pantic (2010)). There is actually no clear consensus in the literature regarding the best length of temporal window to use for a given modality and emotion. Whereas the overall duration of an emotion is supposed to fall between 0.5 and 4 seconds (Levenson (1988)), the length of the analysis window used for emotion prediction can vary greatly according to the modality; audio signals usually change more rapidly over time than video signals, and even more than physiological signals (Kim (2007); Gunes and Pantic (2010); Gunes and Schuller (2013)). In this study, we chose to use various window sizes (ranging from 0.48 s to 6.24 s.) to perform emotion prediction from different modalities (i. e., audio, video and physiological data), and thus estimate which window size performs best for which modality and emotion.

Next to the challenge of time-continuous labels, there is the challenge of multimodal fusion (to improve prediction accuracy) and synchronisation of various individual ratings (to define a reliable gold standard). Regarding multimodal fusion, two main approaches are used in the literature: feature-level and decision-level fusion. Feature-level fusion is performed by merging all the features from multiple modalities into one feature vector (streams with different frame rate can be down-sampled or up-sampled to a common frame rate), which is then given to a machine learning algorithm (Nicolaou et al. (2011); Metallinou et al. (2011)). Whereas in the decision-level fusion, each modality is processed separately by a first emotion recogniser, and another model is trained on the unimodal predictions to predict the actual single-modal gold standard (Kanluan et al. (2008); Nicolaou et al. (2010a)).

Finally, the issue of synchronisation of various individual ratings for defining a gold standard has been investigated using signal processing techniques in the literature. Models of reaction lag have been estimated from the data, by maximising the correlation coefficient (Nicolaou et al. (2010b); Nicolle et al. (2012); Mariooryad and Busso (2013)), or the mutual information (Mariooryad and Busso (2014)) between audiovisual features and emotional ratings while shifting back in time the latter. Such models thus make it possible to compensate for lags when averaging over ratings to obtain a single gold standard. In this study, we chose an approach based on machine learning that is capable of modelling time series with contex-

Table 1. Partitioning of the RECOLA database into train, dev(elopment), and test sets for continuous emotion recognition.

#	train	dev	test
female	5	5	5
male	4	4	4
French	7	7	6
Italian	1	1	3
German	1	1	0
age μ (σ)	20.6 (1.7)	22.6 (4.2)	21.6 (1.1)

tual dependencies: (Bidirectional) Long Short-Term Memory ((B)LSTM) Recurrent Neural Networks (RNN). This type of network has been successfully used in previous work on dimensional emotion recognition (Eyben et al. (2010a, 2012)), and is well suited for working with long time delays (up to multiple seconds) between labels and inputs, such as those arising from de-synchronised modalities and ratings.

3. Data and Methods

This section describes the data and methods used. The new RECOLA multimodal database employed in this study is introduced in the following section. Methods employed to extract features from audio, video and physiological recordings are then described, followed by a description of the LSTM framework that is used to perform automatic emotion prediction on time-continuous ratings.

3.1. The RECOLA multimodal database

A new multimodal corpus of spontaneous interactions in French called RECOLA, for REmote COLlaborative and Affective interactions, was recently introduced by Ringeval et al. (2013). Spontaneous interactions were collected during the resolving of a collaborative task ("Winter survival task", Hall and Watson (1970)) that was performed in dyads and remotely through video conference. The RECOLA database includes 9.5 h of multimodal recordings, i. e., audio, video, electrocardiogram (ECG) and electro-dermal activity (EDA), that were continuously and synchronously recorded from 46 participants. In addition to these recordings, emotional ratings were performed by six French-speaking assistants via the ANNEMO web-based annotation tool, i. e., time- and value-continuous, for the first five minutes of all recorded sequences. The dataset for which participants gave their consent to share their data is reduced to a set of 34 participants for an overall duration of 7 hours, including 5.5 hours of fully multimodal recordings from 27 participants, due to issues during the recording of the physiological signals. The RECOLA database and the ANNEMO web-based annotation tool are both publicly available[1].

For the purpose of this study, we used the 27 subjects of the RECOLA database for which all three modalities were available (audio, video and physiological recordings). Data were divided into speaker disjoint subsets for training, validation and testing, by stratifying (balancing) on gender and mother tongue, cf. Table 1. Regarding the time-continuous annotations of emotion, we used a new normalisation technique to increase the inter-rater agreement, while preserving the original balancing of the ratings: we weight the mean of each rater by his/her respective agreement with others – similar to the Evaluator Weighted Estimator (Grimm et al. (2008)), when defining the reference point used for normalisation of the ratings. We used as metric of inter-rater agreement the mean pair-wise correlation coefficient: the agreement of a rater is obtained by averaging the correlation coefficient over all pairs of raters formed with this rater – $N-1$ pairs for N raters. In order to compare the influence of this technique between raw and zero-mean normalisation, which is the common approach used in the literature, we computed the percentage of positive instances and the concordance correlation coefficient (ρ_c) (cf. Li (1989)), which combines the Pearson correlation coefficient (ρ) and the mean square error (MSE) in a single metric:

Table 2. Influence of different rating normalisation techniques (raw: no normalisation, zero-m.: zero-mean normalisation, wgt-m.: rater agreement weighted-mean normalisation) applied on ratings of arousal and valence provided by 6 annotators; % pos: percentage of positive instances.

	Arousal	**Valence**
raw ρ_c	0.28	0.37
raw % pos	59.0	70.5
zero-m. ρ_c	0.33	0.43
zero-m. % pos	50.8	44.8
wgt-m. ρ_c	0.33	0.43
wgt-m. % pos	48.5	74.1

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

where ρ is the Pearson correlation coefficient between two raters, σ_x^2 and σ_y^2 the variance of each rater, and μ_x and μ_y the mean value of each rater.

Results show that the same amount of improvement can be obtained on the inter-rater agreement (measured as average ρ_c between all rater pairs) when using either zero-mean or weighted-mean normalisation, with a better conservation of the original balance (% pos) for the latter technique, cf. Table 2.

Because we used windows of different size to perform time-continuous emotion prediction, we evaluated the information loss when a mean sliding window is applied on the original ratings (25 Hz). For each rater, we filtered the data with a mean sliding window (shifted forward at a constant rate of 0.48 s) and then interpolated the output to get back to the original 25 Hz frame rate. The result of this interpolation was compared with the original rating by computing the ρ_c. Results show that the information loss increases more rapidly for arousal than for valence when the length of the sliding window increases, since the emotional valence changes less rapidly over time than the arousal (Levenson (1988)). Further, a window up to 6 seconds can be used for emotion prediction with an information loss of

[1] http://diuf.unifr.ch/diva/recola

Table 3. ComParE acoustic feature set: 65 low-level descriptors (LLD).

4 energy related LLD	**Group**
Sum of auditory spectrum (loudness)	prosodic
Sum of RASTA-filtered auditory spectrum	prosodic
RMS Energy, Zero-Crossing Rate	prosodic
55 spectral LLD	**Group**
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	spectral
MFCC 1–14	cepstral
Spectral energy 250–650 Hz, 1 k–4 kHz	spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	spectral
Spectral Flux, Centroid, Entropy, Slope	spectral
Psychoacoustic Sharpness, Harmonicity	spectral
Spectral Variance, Skewness, Kurtosis	spectral
6 voicing related LLD	**Group**
F_0 (SHS & Viterbi smoothing)	prosodic
Prob. of voicing	voice qual.
log. HNR, Jitter (local & δ), Shimmer (local)	voice qual.

less than 0.3 ρ_c for arousal; 0.2 ρ_c for valence.

3.2. Multimodal feature extraction

In order to extract relevant information from signals of audio, video and physiological modalities for emotion prediction, we perform different processing steps on these signals. The goal is to reduce the quantity of data given to the machine learning algorithm, while increasing the relevance for the prediction of emotion.

3.2.1. Audio features

For the extraction of acoustic features, we chose the same set of acoustic Low-Level Descriptors (LLD) as in the last two INTERSPEECH Computational Paralinguistics ChallengEs (ComParE 2013–2014) (Schuller et al. (2013, 2014)). Our open-source feature extractor openSMILE (Eyben et al. (2010b)) was used for this purpose in its recent 2.0 release (Eyben et al. (2013)). The ComParE feature set contains 65 LLD of speech with their first order derivate – 130 LLD in total. Voicing related LLD are extracted from 60 ms frames (Gaussian window function) with $\sigma = 0.4$, all other LLD are extracted from 25 ms frames (Hamming window function). All windows are overlapping and are sampled at a common rate of 100 Hz (10 ms period). For details on the feature set including an in-depth analysis of the features for speech tasks the reader is referred to Weninger et al. (2013). The LLD included in the set are summarised in Table 3.

3.2.2. Video features

As visual features we extracted 20 LLD and their first order derivate (40 LLD in total) for each frame (25 Hz) in the video recordings: 15 facial actions units (AU) involved in emotional expressions, head-pose in three dimensions and the mean and standard deviation of the optical flow in the region around the head. The initial step of visual features extraction is to automatically detect the face region. A facetracker based on the supervised descent method (SDM) was used for this purpose (Xiong and De la Torre (2013)). It estimates the location and the shape of the face using a cascade of regression models that are learned from local texture features (SIFT model); 49 landmarks of the face are returned by the facetracker for each frame, as well as a probability measure of correct detection (estimated by an SVM trained with face and non-face images). LLD of the frames for which the tracker was unable to locate the landmarks, e. g., when the face was occluded or outside, were set to 0. The optical flow was computed around the head region using Farneback's algorithm (Farnebäck (2003)), then the mean and the variance of the norm of the optical flow vectors were calculated. The region of interest was defined by the extremum landmarks from the tracked face shape. This region was enlarged with a safety margin equally on all sides, such that the width and the height are twice of those computed from the tracked face. The 3D head pose angles were estimated in an iterative manner using the tracked (2D) face shape and a 3D point distribution model (PDM), cf. Chen et al. (2012). The AUs, which quantify the muscle activity on the face according to the facial action coding system (FACS, Ekman and Friesen (1978)), were detected by using a linear SVM with SIFT features (Lowe (2004)). For each tracked face image we first apply a texture alignment using a 3D Cylindrical Head Model (CHM, Xiao et al. (2003)) to avoid the effects of varying head-pose. The aligned face image is then scaled to a fixed size of 200 by 200 pixels, and SIFT descriptors are extracted in a 32 x 32 local neighbourhood around each of the 49 landmarks; PCA is applied to reduce dimensionality - 98% of the total variance is retained in the training set. A linear SVM was finally trained with a 5-fold cross validation on the CK+ dataset to detect each AU; distance to the hyperplane was computed as a measure of intensity. Figure 1 shows an example of the face tracking and optical flow calculation result. The detection accuracies for a leave-one-subject-out test on CK+ for the 15 AUs (as well as their definition) are shown in Table 4.

Table 4. AU detection accuracy on the CK+ database; OA(%): Overall accuracy; AUC(%): Area under ROC curve.

AU	OA	AUC
1 (Inner brow raiser)	90.14	93.08
2 (Outer brow raiser)	90.75	91.54
4 (Brow lowerer)	84.16	90.17
5 (Upper lid raiser)	92.08	93.50
6 (Cheek raiser)	85.02	87.32
7 (Lid tightener)	82.09	83.50
9 (Nose wrinkler)	96.25	98.21
11 (Nasolabial deepener)	93.54	79.76
12 (Lip corner puller)	94.16	95.22
15 (Lip corner depressor)	92.14	92.26
17 (Chin raiser)	89.96	95.24
20 (Lip strecher)	96.12	96.76
23 (Lip tightener)	90.91	89.22
24 (Lip pressor)	92.56	87.72
25 (Lips part)	93.19	97.37

3.2.3. Physiological features

Because of their physiological nature, bio-potential signals such as ECG and EDA require more time than audio and video

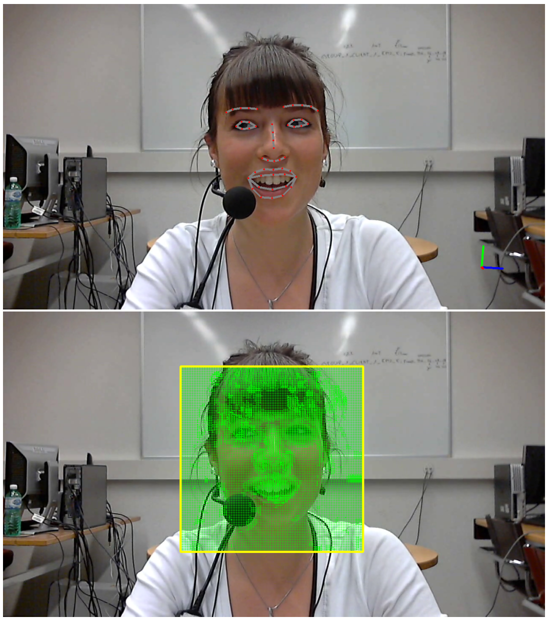

Fig. 1. Top: face tracking and head-pose estimation on a video frame; Bottom: optical flow calculated around the head region.

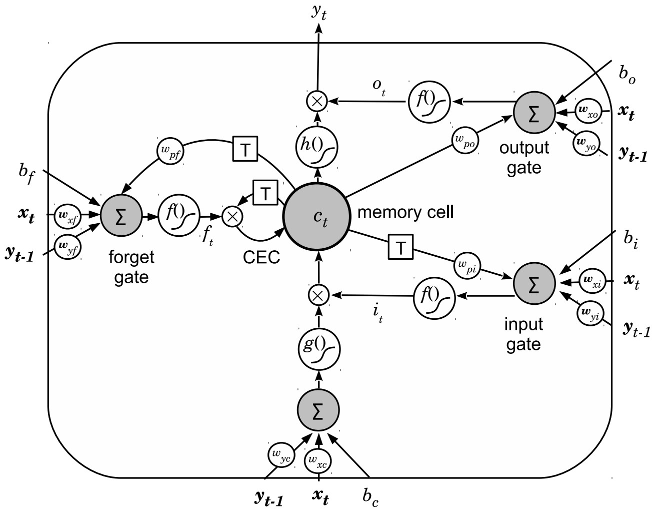

Fig. 2. LSTM block, containing a memory cell and the input (i), output (o) and forget (f) gates. State shown at timestep t. Input data vector (x), connection weights w_{ab} (multiplicative), bias values b, cell output y. Nonlinear squashing functions $g()$ and $h()$. The vector containing all outputs of the current hidden layer at timestep t is denoted as y_t. T denoting a time delay unit of one timestep. X in a circle denoting a multiplicative unit. Σ denotes a summation unit. $f()$, $g()$, and $h()$ are activation functions (nonlinearities). (Eyben, 2014)

signals for extracting LLD. In this study, LLD were extracted from both ECG and EDA signals with overlapping (step of 0.48 s) windows of 4 s length. 28 LLD were extracted in total from the ECG signals: the heart rate (HR) and its measure of variability (HRV), the zero-crossing rate, the 4 first statistical moments (Picard et al. (2001)), the normalised length density (NLD) and the non-stationary index (NSI), the spectral entropy, slope and mean frequency plus 12 spectral coefficients (Hanning overlapping windows ranging from 3 to 27 Hz), and the power of HR in low frequency (LF, 0.04-0.15 Hz), high frequency (HF, 0.15-0.4 Hz) and the LF/HF ratio (Bilchick and Berger (2006)); the first order derivate was furthermore computed on all excepted HR and HRV, which thus provided 54 LLD in total for the ECG signals.

In order to extract the heart rate (HR), the interval between two QRS complexes defined as R-R interval (t_{R-R}) was estimated using the real-time algorithm developed by Pan and Tompkins (1985). The respiration drift was removed using a morphological operator (maxima and minima computed on a sliding window). The NSI is a measure of signal complexity; it segments the signals into small parts and estimates the variation of the local averages (Hausdorff et al. (2000)). Finally, the NLD index was extracted to capture non-linear temporal variations of the signals:

$$\mathrm{NLD} = \frac{1}{N} \sum_{i=2}^{N} |\mathrm{y_n(i)} - \mathrm{y_n(i-1)}|, \qquad (2)$$

where $\mathrm{y_n(i)}$ and N represent the ith sample after amplitude normalization and the length of the signal respectively (Kalauzi et al. (2009)).

EDA reflects a rapid, transient response called skin conductance response (SCR), as well as a slower, basal drift called skin conductance level (SCL) (Dawson et al. (2007)). Both SCL (0–0.5 Hz) and SCR (0.5–1 Hz) were extracted using a 3rd order Butterworth filter, 30 LLD were then computed in total: the temporal slope of EDA (first coefficient of a first order regression polynomial), the spectral entropy and mean frequency of SCR, the NSI and NLD index, the 4 first statistical moments, the mean value of the first order derivate, and the proportion and mean of its negative part for EDA, SCL and SCR. Finally, first order derivate of these LLD were computed, which provided 60 LLD in total for EDA signals.

3.3. Continuous emotion prediction

3.3.1. Memory-enhanced networks

Traditional Feed-Forward Neural Networks (FF-NN) with sigmoid summation units have no memory or feedback connections, i. e., they have no knowledge about other inputs/frames than the current time step. A logical extension is to make the network recurrent, i. e., add a feedback from the output to the inputs with a delay of one timestep. Such networks are known as RNN. However, these networks suffer from the 'Vanishing Gradient Problem' (Hochreiter et al. (2001)), where the activations and/or error on the recurrent connection decay exponentially. This limits the amount of temporal context that is accessible to the networks to approximately 10 frames. To overcome this problem, LSTM-RNN have been introduced originally by Hochreiter and Schmidhuber (1997), and extended to the version used in this article by Graves and Schmidhuber (2005). The main difference between the original version and the version used in this article is the use of *peep-hole* connections from the internal memory state to the input, output, and forget gate summation units (cf. w_{px} in Figure 2). The sigmoid summation units in the hidden layers of a conventional RNN are replaced by so-called LSTM memory blocks in LSTM-RNN. These LSTM blocks can store information in the cell variable

c_t for an indefinite amount of time due to the Constant Error Carousel (CEC) where the previous memory cell state c_{t-1} is connected to the current state over a recurrent connection with constant weight 1 (without the dynamic multiplicative influence of the forget gate), cf. Figure 2. In this way, the network can dynamically exploit long-range temporal context without the Vanishing Gradient problem.

Each LSTM block contains a memory cell and three multiplicative gates: the input gate, the output gate and the forget gate (Figure 2). These gates guard the data-flow to and from the block's internal memory cell c_t. For the input gate, for example, the activation at the output of the gate is computed as:

$$i_t = f(\boldsymbol{w}_{xi}\boldsymbol{x}_t + \boldsymbol{w}_{yi}\boldsymbol{y}_{t-1} + w_{pi}c_{t-1} + b_i), \tag{3}$$

where $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ are weight vectors (row vectors) matching the dimensionality of $\boldsymbol{x}$ or $\boldsymbol{y}$, respectively. $\boldsymbol{x}_t$ is the input vector at time step t, $\boldsymbol{y}_{t-1}$ is the hidden state vector of the previous time step, and b_i denotes the input gate bias value for this cell. The activations of the forget and output gates (f_t and o_t) are computed in the same way. The computations can be derived from Figure 2. The forget gate controls the retention or decay of the stored input c_t. If $f_t = 0$, the previous cell state c_{t-1} is fully deleted. The input and output gates are responsible for dynamically weighting the cell input and output, respectively. The internal cell state c_t at timestep t is given as:

$$c_t = f_t c_{t-1} + i_t g(\boldsymbol{w}_{xc}\boldsymbol{x}_t + \boldsymbol{w}_{yc}\boldsymbol{y}_{t-1} + b_c), \tag{4}$$

and the output activation of the cell is:

$$\boldsymbol{y}_t = o_t h(c_t). \tag{5}$$

The functions f (for the gates), g (for the input), and h (for the output) are non-linear squashing functions, just as those in normal sigmoid neural network units. The sigmoid (logistic) function is used for the gate activation (function f) and the *tanh* function is used for the cell inputs and outputs (function g).

In addition to LSTM memory blocks as hidden units, we employ bidirectional LSTM-RNN (Schuster and Paliwal (1997) (BLSTM-RNN)). A bidirectional RNN theoretically has access to all past and all future inputs, which renders it ideal for processing data with de-synchronisation between inputs and targets. This property is made possible by processing the data in both directions in two separate hidden layers: one processes the data sequence forwards, the other one backwards. 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-RNN (Graves and Schmidhuber (2005)).

3.3.2. Multi-task learning

We consider two types of multi-task learning: one by learning each rater's individual track, i. e., six raters' values are learnt and output per dimension, and then averaged at the output as compared to only learning and directly outputting the mean, as is 'traditionally' done. In addition to this, we also consider the learning of both dimensions simultaneously (and potentially also from several raters simultaneously). Any of the multi-task learning is simply realised by adding accordingly more output nodes to the networks.

4. Experiments and Results

In this article, we contrast multi-task and single-task LSTM-RNN, BLSTM-RNN and as a 'non-context aware baseline' FF-NN. We perform experiments on the RECOLA database for the dimensions arousal and valence. Since a recent study has shown that a duration of less than 0.5 s seems to be suitable to define micro-expressions in leaked facial expressions (Yan et al. (2013)), and the analysis of the time-continuous ratings has shown that a window up to 6 s can be used to describe changes in emotion without losing too much information, cf. Section 3.1, we used overlapping windows of different sizes, ranging from 0.48 s to 6.24 s with a common shift step of 0.48 s, to compute functionals (min, max, range, mean and standard-deviation) on LLD from all four modalities; because of the low frame rate available on the LLD computed on ECG and EDA signals, only a window size superior or equal to 1.92 s was used for these LLD.

Due to the use of functionals, the number of effective features at the input of the (B)LSTM-RNN and FF-NN is multiplied by 5. For multimodal experiments with a feature-level fusion, all features are simply concatenated frame by frame on the feature level; the window size providing the best performance is retained for each modality regarding both network architecture (i. e., FF-NN, (B)LSTM-RNN) and learning scheme (single-, multi-task on the mean or on all ratings). For the decision-level fusion, we used a linear Support Vector Regression with the complexity value trained on predictions made on the development partition; the combination of window size, learning scheme and network architecture leading to the best performance on the test partition is selected for each modality. Additionally, we include as feature the probability of face detection for the visual modality (provided by the face tracker, cf. Section 3.2.2), and the probability of speech production for the audio modality (segmentation into speech turns is provided in the corpus). The mean probability of face detection is equal to 0.90 (± 0.17) on dev and 0.85 (± 0.18) on test; the mean probability of speech production is equal to 0.39 (± 0.11) on dev and 0.43 (± 0.10) on test.

4.1. Training and setup of the Neural Networks

Before training the neural networks, all inputs and all targets are normalised to zero mean and unit variance, using parameters (mean, variance) computed from the training set. The resulting predictions from the networks are unnormalised (inversion of the mean/variance normalisation) before they are compared to the ground truth in evaluations. For network training, we use batch gradient descent by backpropagation through time (Werbos (1990)), where weight changes are applied after processing all training sequences at the end of each training epoch (Graves (2008)). There, the weights are updated after every sequence. Here, in order to run the error computation in parallel for a given number of sequences (for computational speed-up), we use mini batches of 10 sequences. The training is performed with the CURRENNT toolkit[2]. To improve generalisation and prevent overfitting, Gaussian noise with a standard

[2] `http://sourceforge.net/projects/currennt/`

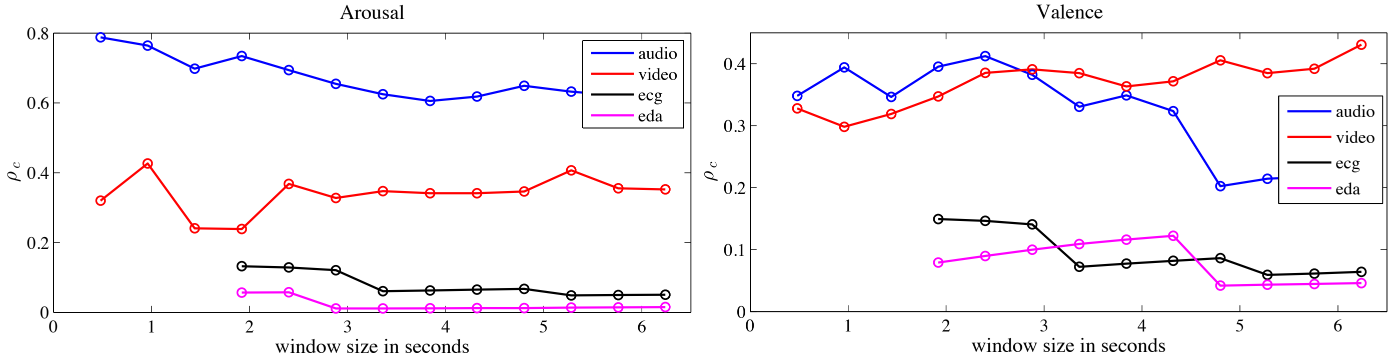

Fig. 3. Evolution of performance (in terms of ρ_c) in the automatic prediction of (left: arousal, right: valence) for audio, ECG, EDA and video modalities over various window sizes; results are shown for the combination of network architecture and learning scheme that led to the best performance (maximum of ρ_c) according to each modality; only a window size superior or equal to 1.92 s was used for ECG and EDA, because of the low frame rate of their LLD.

deviation 0.1 is added to all input features. The networks are trained for a maximum of 100 epochs. Training is stopped if no improvement of the performance by MSE is observed on the development set for more than 20 epochs. The best network (on the development set) is chosen and the performance on the test set is evaluated. The BLSTM networks used have two hidden layers with 40 and 30 BLSTM blocks, respectively. The (unidirectional) LSTM networks have 80 and 60 LSTM blocks, respectively, in order to keep the amount of parameters in LSTM and BLSTM the same. Each layer is fully connected. Recurrent connections are only from a layer's output to its input, i. e., recurrent connections do not go back more than one layer. The output layer consists of linear summation units (no sigmoid squashing function). The number of units depends on the number of outputs. For single task networks (arousal or valence) there is only one output, while for multi-task networks there are two outputs. For networks which estimate each rater individually, there are 6 or 12 outputs (single- and multi-task, respectively). For comparison with non-context aware learning methods, FF-NN with 160 sigmoid units in the first hidden layer and 120 sigmoid units in the second hidden layer are implemented.

4.2. *Window size vs. modality & emotion*

For each modality (audio, ECG, EDA and video) and dimension (arousal and valence), we selected the best configuration of the learning scheme (single-, multi-task learning, on the mean rating or on all ratings) and network architecture (i. e., FF-NN, (B)LSTM-RNN) according to the performance (ρ_c) obtained on our test partition. The evolution of performance over the window size for these best configurations is shown for each modality and emotion in Fig. 3. Results show that features computed for the audio modality provide the best performance for the prediction of arousal, whereas video performs best for valence. Moreover, the automatic prediction of the emotion performs much better on arousal than on valence. These results are in agreement with the literature (Gunes and Pantic (2010); Gunes and Schuller (2013)).

Despite the information loss on emotion increases along with window size (cf. Section 3.1), noise in the time-continuous ratings (e. g., raters may not be sure of their feedback and thus move the cursor around the depicted emotion), can be better filtered out when using longer analysis windows, especially for valence, which is a more subjective emotion than arousal; Cronbach's α on the 6 ratings of arousal is 0.80, whereas it is 0.74 for valence; cf. Table 3 in Ringeval et al. (2013). The performance obtained on the ECG and EDA signals is however quite low on both arousal and valence ($\rho_c < 0.2$). A possible explanation might be the fact that subjects were moving considerably during the experiments, because they were allowed to take notes during the discussions. Thus, possible noise due to those movements might have decreased the performance of the physiological modality. Moreover, ratings of arousal and valence were performed only with audio and video modalities available.

Regarding the best window size for the prediction of emotion, results show that valence requires in the mean (over the 4 modalities) a window about twice more the duration of the one used for arousal to achieve best performance; the mean window length over the 4 modalities is 1.44 s (± 0.88) for arousal and 3.72 s (± 1.97) for valence. This result is coherent with the findings of the analysis performed on the ratings - arousal changes more rapidly over time than valence - cf. Section 3.1.

4.3. *Mean ratings vs. all ratings*

We compare the performance on ρ_c for each 4 modalities (according to their respective best network and window size) between mean ratings (average over all raters) and all 6 ratings, for each type of learning task (i. e., arousal and valence separately or all together) in Table 5. Statistical tests (1-way ANOVA) were performed between the two types of prediction (i. e., all ratings or mean ratings) on all frames, i. e., more than 10k instances were compared in total. Results show that performance can be significantly improved when using all 6 ratings in a single network (network with 6 outputs), compared with the use of mean ratings, for predictions of arousal and valence with video data. This lets us suppose that the network can deal with (asynchronous) dependencies found between all 6 raters, and even make use of the individual information given by each rater. However, this observation cannot be made on audio data; since performance was found significantly better on mean ratings than for all ratings. A more complex network (more hid-

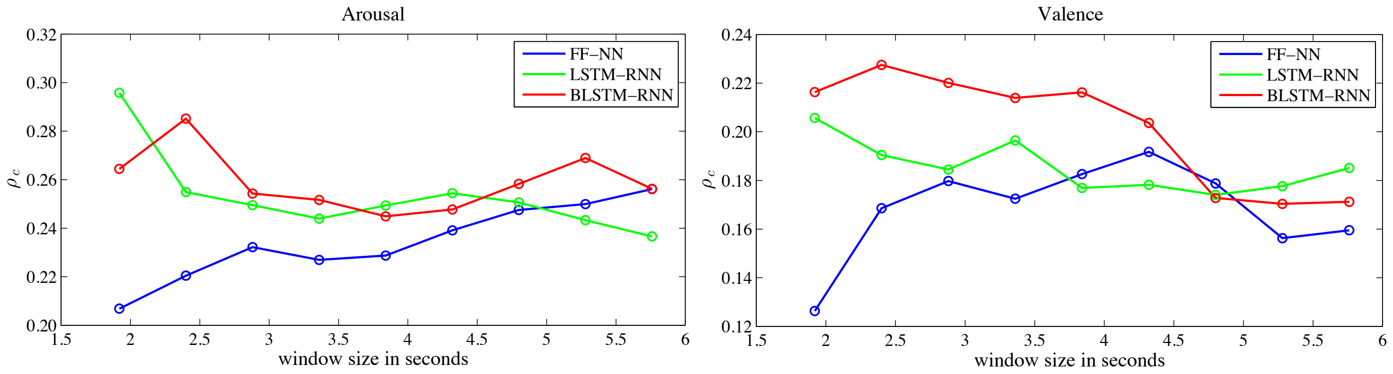

Fig. 4. Evolution of performance (in terms of ρ_c) in the automatic prediction of (left: arousal, right: valence) over various window sizes for different types of neural network architecture; performance is averaged over all 4 modalities (audio, ECG, EDA and video) according to their respective best learning scheme (single-, multi-task learning, on mean or all ratings).

Table 5. Performance obtained on single- and multi-task prediction of *mean* ratings – or *all* six ratings – of arousal and valence; performance (ρ_c) is computed for each modality according to their respective best network and window size; a * indicates that the predictions differ significantly ($p < 0.05$) between mean and all.

ρ_c	**single task**		**multi task**	
	mean	all	mean	all
		AROUSAL		
Audio	**0.788***	0.757	0.732	**0.738**
ECG	0.062	0.052	0.132	0.033
EDA	0.057	0.051	0.053	0.049
Video	0.390	0.382	0.403	0.427*
		VALENCE		
Audio	0.292*	0.260	0.412*	0.343
ECG	0.097	0.005	0.149	0.037
EDA	0.109	0.106	0.111	0.122
Video	0.409	**0.431***	0.339	**0.349**

den units) might be needed, because the complexity of the task is higher for the prediction of all 6 ratings (the number of outputs is multiplied by 6), especially for audio data which contain 3 times more LLD than for video data.

4.4. Comparison of Neural Networks

We show in Fig. 4 the average of performance on ρ_c (over all 4 modalities) obtained by each neural network, i. e., FF-NN, LSTM-RNN and BLSTM-RNN, when predicting arousal or valence; best configuration (i. e., single-, multi-task learning on mean or all ratings) was selected for each combination of modality and network. Results show that the performance obtained with FF-NN increases with the window size for both arousal and valence (a drop is observed on valence after 4.5 s), whereas the performance obtained with LSTM-RNN and BLSTM-RNN drops for longer windows. This effect, which is statistically significant (2-way ANOVA on the 9 measures of performance; $p < 0.05$), shows that the LSTM networks can deal with short segments and assemble the context through their memory cells from the neighbouring frames, while the FF-NN requires the context to be present in the features through the use of longer analysis windows.

4.5. Comparison of multimodal fusion approaches

We performed multi-modal fusion for all combinations of modalities by using either a feature-level or a decision-level fusion. Additionally, we include as feature the probability of face detection for the visual modality, and the probability of speech production for the audio modality. Results show that a decision-level fusion provides best performance for both arousal and valence, cf. Table 6. The emotion information conveyed by each modality is thus better modelled when specific machine learning algorithms are separately used for each modality. The use of the knowledge of speech production and face detection as feature does not improve the performance, which lets us suppose that the system already learns by itself which modality to trust over time. Finally, the combination of modalities leading to the best performance for the prediction of arousal includes audio, video and EDA, and all 4 modalities (i. e., including also ECG) for the prediction of valence, which thus shows the interest of using both audiovisual and physiological information to perform emotion recognition on arousal (especially EDA) and valence (both ECG and EDA). The difference between the gold standard and the automatic prediction obtained with the best setting is shown frame by frame for a single test subject for both arousal and valence in Figure 5.

5. Conclusion

We investigated in this paper the relevance of using machine learning algorithms able to integrate contextual information in the modelling, like the employed LSTM-RNN do, in order to automatically predict emotion from several (asynchronous) raters in continuous time domains, i. e., arousal and valence. Evaluations were performed on the recently proposed RECOLA multimodal database, with both mono-modal, i. e., audio, video, or physiology (ECG and EDA) based features, and multimodal approaches, i. e., the fusion of these modalities. Automatic emotion prediction performance was evaluated by using different window sizes for features extraction, and different architectures of Neural Networks: one able to integrate

Table 6. Performance (ρ_c) obtained by each combination of modality for feature-level and decision-level fusion, and without or with the inclusion of event based probability for audio (speech production) and video (face detection) data as additional feature; (a): audio modality; (v): video modality.

ρ_c	**Feature-level**		**Decision-level**	
	without	with	without	with
AROUSAL				
a+ecg	0.681	0.721	0.783	0.784
a+eda	0.696	0.693	0.793	0.794
a+v	**0.761**	**0.769**	0.796	0.796
ecg+eda	0.063	0.063	0.145	0.145
ecg+v	0.361	0.346	0.409	0.452
eda+v	0.366	0.397	0.409	0.441
a+ecg+eda	0.627	0.604	0.790	0.791
a+ecg+v	0.737	0.741	0.796	0.797
a+eda+v	0.700	0.690	**0.804**	**0.804**
ecg+eda+v	0.284	0.290	0.416	0.450
a+ecg+eda+v	0.671	0.691	0.802	0.803
VALENCE				
a+ecg	0.309	0.342	0.368	0.375
a+eda	0.276	0.349	0.424	0.444
a+v	**0.492**	0.443	0.501	0.492
ecg+eda	0.085	0.085	0.095	0.095
ecg+v	0.380	0.304	0.325	0.335
eda+v	0.287	0.277	0.330	0.345
a+ecg+eda	0.307	0.347	0.445	0.453
a+ecg+v	0.432	**0.472**	0.490	0.482
a+eda+v	0.414	0.446	0.525	**0.527**
ecg+eda+v	0.249	0.253	0.334	0.341
a+ecg+eda+v	0.302	0.351	**0.528**	0.523

contextual information (i. e., LSTM-RNN), and another that does not include such information (i. e., FF-NN). The results showed that the prediction of the emotional valence requires longer analysis window than for arousal (about twice more the duration), and that the integration of contextual information in the modelling of emotion is needed in order to include reaction time delay of raters. Moreover, performance can be significantly improved when using all ratings in a single network, for the prediction of both arousal and valence on video data, which lets us suppose that the network can deal with dependencies found between all available raters (6), and even make use of the individual information given by each rater. However, this was not observed on audio data; more complex networks (more hidden units) might be needed, in order to cope with the increase of complexity of the task for the prediction of all 6 ratings with a large number of LLD (i. e., > 100). Finally, a decision-level fusion provided a better performance than a feature-level fusion on the prediction of both arousal and valence, showing the complementarity of audiovisual and physiological data for emotion recognition, especially EDA for arousal and both ECG and EDA for valence; the best performance (ρ_c) for the multimodal emotion prediction is 0.804 for arousal and 0.528 for valence.

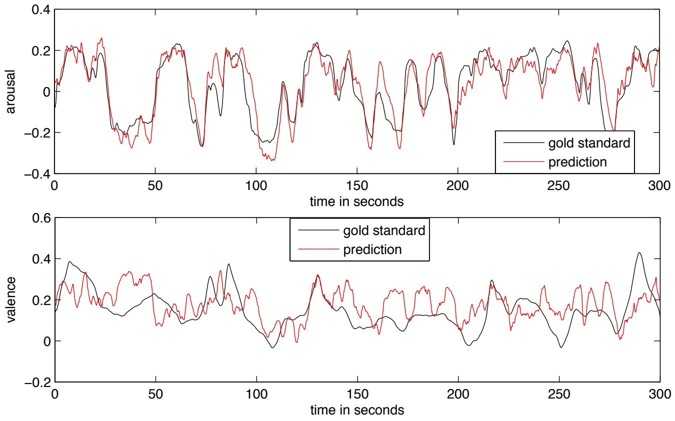

Fig. 5. Automatic prediction of arousal (top) and valence (bottom) obtained with the best setting for a subject from the test partition.

Because SVM and SVR are successful in utterance level emotion recognition, recurrent SVM/SVR (Schneegaß et al. (2007)) could be investigated in future work. Inter-rater synchronisation, as well as synchronisation between features and raters could also be investigated, using correlation-based analysis methods, to compare the impact of using contextual information independently of lag in the ratings, on emotion recognition performance.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreements No. 338164 (ERC Starting Grant iHEARu), and has been supported by the Swiss National Science Foundation through the National Centre for Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

References

Bilchick, K.C., Berger, R.D., 2006. Heart rate variability. Journal of Cardiovascular Electrophysiology 17, 691–694.

Chen, S.C., Wu, C.H., Lin, S.Y., Hung, Y.P., 2012. 2D face alignment and pose estimation based on 3D facial models, in: Proc. of the 12th Inter. Conf. on Multimedia & Expo (ICME), IEEE, Melbourne, Australia. pp. 128–133.

Chetouani, M., Mahdhaoui, A., Ringeval, F., 2009. Time-scale feature extractions for emotional speech characterisation. Cognitive Computation 1, 194–201.

Dawson, M., Schell, A., Filion, D., 2007. The electrodermal system, in: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (Eds.), Handbook of psychophysiology. Cambridge: Cambridge University Press. volume 2, pp. 200–223.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.C., Devillers, L., Abrilian, S., Batliner, A., Amir, N., Karpouzis, K., 2007. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data, in: Paiva, A., Prada, R., Picard, R. (Eds.), 2nd Inter. Conf. on Affective Computing and Intelligent Interaction (ACII), LNCS. Springer-Verlag Berlin Heidelberg, Lisbon, Portugal. volume 4738, pp. 488–500.

Ekman, P., Friesen, W.V., 1978. Facial action coding system: A technique for the measurement of facial movement. Palo Alto, CA: Consulting Psychologists Press.

Eyben, F., 2014. Real-time speech and music classification by large audio feature space extraction. Ph.D. thesis. Technische Universität München. Submitted, to appear.

Eyben, F., Weninger, F., Groß, F., Schuller, B., 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor, in: Proc. of the 21st ACM Inter. Conf. on Multimedia (ACM MM), Barcelona, Spain. pp. 835–838.

Eyben, F., Wöllmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., Cowie, R., 2010a. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. Journal on Multimodal User Interfaces 3, 7–12.

Eyben, F., Wöllmer, M., Schuller, B., 2010b. openSMILE – The Munich versatile and fast open-source audio feature extractor, in: Proc. of the 18th ACM Inter. Conf. on Multimedia (ACM MM), Florence, Italy. pp. 1459–1462.

Eyben, F., Wöllmer, M., Schuller, B., 2012. A multi-task approach to continuous five-dimensional affect sensing in natural speech. ACM Trans. on Interactive Intelligent Systems (TiiS) - Special Issue on Affective Interaction in Natural Environments 2. 29 pages.

Farnebäck, G., 2003. Two-frame motion estimation based on polynomial expansion, in: Image Analysis, Special Issue, 13th Scandinavian Conf. (SCIA). Springer, Halmstad, Sweden. volume 2749, pp. 363–370.

Graves, A., 2008. Supervised sequence labelling with recurrent neural networks. Ph.D. thesis. Technische Universität München. Munich, Germany.

Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, Special Issue, 18th Inter. Joint Conf. on Neural Networks (IJCNN) 18, 602–610.

Grimm, M., Kroschel, K., Narayanan, S., 2007. Support vector regression for automatic recognition of spontaneous emotions in speech, in: Proc. of the 32th IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, HI, USA. pp. 1085–1088.

Grimm, M., Kroschel, K., Narayanan, S., 2008. The Vera am Mittag German audio-visual emotional speech database, in: Proc. of the 8th Inter. Conf. on Multimedia & Expo (ICME), Hannover, Germany. pp. 865–868.

Gunes, H., Pantic, M., 2010. Automatic, dimensional and continuous emotion recognition. Inter. Journal of Synthetic Emotions 1, 68–90.

Gunes, H., Schuller, B., 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and Vision Computing: Affect Analysis in Continuous Input 31, 120–136.

Hall, J., Watson, W., 1970. The effects of a normative intervention on group decision-making performance. Human Relations 23, 299–317.

Hausdorff, J.M., Lertratanakul, A., Cudkowicz, M.E., Peterson, A., Kaliton, D., Golberger, A.L., 2000. Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. Journal of Applied Physiology 88, 2045–2053.

Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., 2001. Gradient flow in recurrent nets: The difficulty of learning long-term dependencies, in: Kremer, S.C., Kolen, J.F. (Eds.), Field Guide to Dynamical Recurrent Networks. IEEE Press.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9, 1735–1780.

Kalauzi, A., Bojic, T., Rakic, L., 2009. Extracting complexity waveforms from one-dimensional signals. Nonlinear Biomedical Physics 3.

Kanluan, I., Grimm, M., Kroschel, K., 2008. Audio-visual emotion recognition using an emotion recognition space concept, in: Proc. of the 16th European Signal Processing Conf. (EUSIPCO), Lausanne, Switzerland.

Kim, J., 2007. Bimodal emotion recognition using speech and physiological changes, in: Grimm, M., Kroschel, K. (Eds.), Robust speech recognition and understanding. I-Tech Education and Publishing, Vienna, Austria, pp. 265–280.

Levenson, R., 1988. Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity, in: Wagner, H.L. (Ed.), Social psychophysiology and emotion: Theory and clinical applications. John Wiley & Sons, pp. 17–42.

Li, L., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. Inter. Journal of Computer Vision 60, 91–110.

Mariooryad, S., Busso, C., 2013. Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations, in: Proc. of the 5th Inter. Conf. on Affective Computing and Intelligent Interactions (ACII), Geneva, Switzerland. pp. 85–90.

Mariooryad, S., Busso, C., 2014. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. IEEE Trans. on Affective Computing PP, 12.

Metallinou, A., Katsamanis, A., Wang, Y., Narayanan, S., 2011. Tracking changes in continuous emotion states using body language and prosodic cues, in: Proc. of the 36th IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 2288–2291.

Nicolaou, M., Gunes, H., Pantic, M., 2010a. Audio-visual classification and fusion of spontaneous affective data in likelihood space, in: Proc. of the 20th IEEE Int. Conf. on Pattern Recognition (ICPR), Istambul, Turkey. pp. 3695–3699.

Nicolaou, M., Gunes, H., Pantic, M., 2010b. Automatic segmentation of spontaneous data using dimensional labels from multiple coders, in: Proc. of the LREC 2010 Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing, Istambul, Turkey. pp. 43–48.

Nicolaou, M., Gunes, H., Pantic, M., 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Trans. on Affective Computing 2, 92–105.

Nicolle, J., Rapp, V., Bailly, K., Prevost, L., Chetouani, M., 2012. Robust continuous prediction of human emotions using multiscale dynamic cues, in: Proc. of the 14th ACM Inter. Conf. on Multimodal Interaction (ICMI), Istambul, Turkey. pp. 501–508.

Pan, J., Tompkins, W.J., 1985. A real-time QRS detection algorithm. IEEE Trans. Biomedical Engineering 32, 230–236.

Picard, R., Vyzas, E., Healey, J., 2001. Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Trans. on Pattern Analysis and Machine Intelligence 23, 1175–1191.

Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D., 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions, in: Proc. of Face & Gestures 2013, 2nd IEEE Inter. Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE), Shanghai, China.

Schneegaß, D., Schaefer, A.M., Martinetz, T., 2007. The intrinsic recurrent support vector machine, in: Proc. of the 15th IEEE European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium. pp. 325–330.

Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., de Sevin, E., Valstar, M., Wöllmer, M., 2012. Building Autonomous Sensitive Artificial Listeners. IEEE Trans. on Affective Computing 3, 165–183.

Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., Zhang, Y., 2014. The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load, in: Proc. of INTERSPEECH 2014, 15th Annual Conf. of the Inter. Speech Communication Association (ISCA), Singapore, Republic of Singapore.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S., 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism, in: Proc. of INTERSPEECH 2013, 14th Annual Conf. of the Inter. Speech Communication Association (ISCA), Lyon, France. pp. 148–152.

Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE Trans. on Signal Processing 45, 2673–2681.

Tversky, A., 1969. Intransitivity of preferences. Psychological Review 76, 31–48.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: Resources, features, and methods. Speech Communication 48, 1162–1181.

Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G., 2007. Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing, in: Proc. of the 2nd ACM Int. Conf. on Affective Computing and Intelligent Interaction (ACII), Lisbon, Portugal. pp. 139–147.

Weninger, F., Eyben, F., Schuller, B.W., Mortillaro, M., Scherer, K.R., 2013. On the acoustics of emotion in audio: What speech, music and sound have in common. Frontiers in Psychology, Emotion Science, Special Issue on Expression of emotion in music and vocal communication 4, 1–12.

Werbos, P., 1990. Backpropagation through time: What it does and how to do it. Proceedings of the IEEE 78, 1550–1560.

Xiao, J., Moriyama, T., Kanade, T., Cohn, J.F., 2003. Robust full-motion recovery of head by dynamic templates and re-registration techniques. Inter. Journal of Imaging Systems and Technology 13, 85–94.

Xiong, X., De la Torre, F., 2013. Supervised descent method and its applications to face alignment, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Portland (OR), USA. pp. 532–539.

Yan, W.J., Wu, Q., Liang, J., Chen, Y.H., Fu, W., 2013. How fast are the leaked facial expressions: The duration of micro-expressions. Journal of Nonverbal Behavior 37, 217–230.