Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio

Felix Weninger[1], **Fabien Ringeval**[2], **Erik Marchi**[2], **Björn Schuller**[2,3]
[1] MISP/MMK, Technische Universität München, 80290 Munich, Germany
[2] Chair of Complex and Intelligent Systems, University of Passau, 94032 Passau, Germany
[3] Department of Computing, Imperial College London, London SW7 2AZ, UK
felix@weninger.de

Abstract

Continuous dimensional emotion recognition from audio is a sequential regression problem, where the goal is to maximize correlation between sequences of regression outputs and continuous-valued emotion contours, while minimizing the average deviation. As in other domains, deep neural networks trained on simple acoustic features achieve good performance on this task. Yet, the usual squared error objective functions for neural network training do not fully take into account the above-named goal. Hence, in this paper we introduce a technique for the discriminative training of deep neural networks using the concordance correlation coefficient as cost function, which unites both correlation and mean squared error in a single differentiable function. Results on the MediaEval 2013 and AV+EC 2015 Challenge data sets show that the proposed method can significantly improve the evaluation criteria compared to standard mean squared error training, both in the music and speech domains.

1 Introduction

Continuous dimensional emotion recognition from audio is a sequential learning problem that has attracted increasing attention in the past few years [Coutinho and Cangelosi, 2010; Schmidt and Kim, 2011; Metallinou *et al.*, 2013; Soleymani *et al.*, 2013; Wang *et al.*, 2015]. There, sequences of acoustic features have to be mapped to emotion contours in several dimensions that represent the emotion communicated by means of audio, e.g., speech utterances or excerpts of music. Typical emotion dimensions comprise arousal and valence [Russell, 1980], as explored in this study, although other dimensions such as dominance and expectation can be added [Eyben *et al.*, 2012]. Defining the target labels as real-valued mappings from time instants to targets helps capturing the temporal dynamics of emotion, which cannot be assumed to be constant over time [Schmidt and Kim, 2011]. To learn such mappings, deep recurrent neural networks are a promising model [Coutinho and Cangelosi, 2010], as they take into account temporal dependencies in inputs and outputs and can handle correlated features.

Continuous emotion recognition is typically evaluated in terms of the correlation between the learner's outputs and the target values (such as by the correlation or determination coefficient), as well as the average deviation of outputs and targets, such as by the mean linear or mean squared error (MLE/MSE) [Schuller *et al.*, 2012; Jenke *et al.*, 2013]. Since neural networks are usually trained using criteria such as the (root) MSE, this only takes into account the latter while neglecting the former. Further, although it is well known that the minimization of the MSE and the maximization of the (Pearson) correlation coefficient (CC) are equivalent if the outputs and targets are standardized, such standardization cannot be assumed in emotion regression, as the emotional intensity, and hence the mean and variance, is of crucial importance.

Moreover, the CC is insensitive to scaling and shifting, which is problematic for training neural networks with this metric. Imposing a cost function based on CC may lead to an infinite number of local minima with different prediction behavior. In fact, a neural network trained on CC cannot learn the correct scales and offsets from the target values ('gold-standard') because the CC is not sensitive to such variations. We illustrate those issues on different variants of the same time-series in Figure 1. Because CC is insensitive to scaling and shifting, it always provides the same perfect prediction score (CC $= 1.00$) on different versions (shifted and/or scaled) of the same time-series, although such variations are actually far from a perfect prediction. As a result, the CC is not sufficient as evaluation criterion in practice, and additional measures such as mean squared error need to be taken into account.

From the above considerations, we can conclude that firstly, the usual objective functions for neural network regression do not fully match the evaluation criteria used for continuous dimensional emotion recognition, and secondly, the use of CC as objective function cannot lead to satisfying results. To alleviate these problems, we propose to use the concordance correlation coefficient (CCC) [Lin, 1989] as a differentiable objective function that unites both correlation and mean squared error, and can be thought of as a CC that enforces the correct scale and offset of the outputs. As a result, CCC takes into account the effects of shifting and scaling the prediction when computing the performance (cf. Fig. 1), and can thus be used as an objective function to train neural networks for time-continuous prediction tasks.

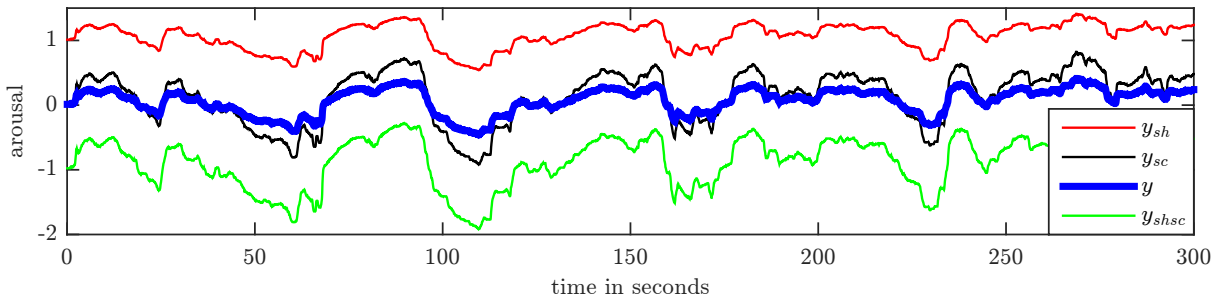

Figure 1: Illustration of the effects of shifting and scaling a gold-standard time-series (arousal) on a subject from the training partition of the RECOLA database) on the Pearson's correlation coefficient (CC) and the concordance correlation coefficient (CCC). The gold-standard y time-series is plotted as a thick blue line; a shifted version ($y_{\text{sh}} = y + 1$) gives CC $= 1$ and CCC $= 0.07$, a scaled version ($y_{\text{sc}} = 2 * y$) provides CC $= 1$ and CCC $= 0.78$, and a shifted and scaled version ($y_{\text{sh,sc}} = 2 * y - 1$) returns CC $= 1$ and CCC $= 0.15$.

In the following, we will show that the choice of objective function (sum of CCCs per sequence, overall CCC, or MSE) for network training significantly influences the evaluation outcome on standard corpora for continuous dimensional emotion recognition from music and speech.

The remainder of this paper is as follows: we describe some related work on task-specific discriminative objective functions for neural network training in section 2, we introduce different discriminative objectives for emotion regression in section 3, and the optimization of CCC by stochastic gradient descent in section 4, we evaluate the performance in dimensional emotion recognition tasks (arousal and valence) for different objective functions on two different corpora (music and speech) in section 5, and provide a conclusion in section 6.

2 Related Work

Task-specific discriminative objective functions for neural network training are well known. For example, in training networks for automatic speech recognition, the minimum phoneme error or minimum Bayes risk objectives are used in place of the standard cross entropy objective [Veselý *et al.*, 2013; Kingsbury *et al.*, 2012]. In [Weninger *et al.*, 2014], it is proposed to optimize the prediction of time-frequency masks for acoustic source separation based on local signal-to-noise ratio rather than MSE. Joint optimization of masking functions and deep recurrent neural networks has been investigated for monaural source separation tasks by [Huang *et al.*, 2015], with a discriminative criterion designed to enhance the source to interference ratio.

Regarding neural network approaches for emotion recognition, most of the existing work has been focused on classification tasks. Deep Neural Network Hidden Markov Models (DNN-HMMs) were investigated with discriminative pre-training and restricted Boltzmann Machine (RBM) based unsupervised pre-training by [Li *et al.*, 2013]. Experimental results have shown the superiority of the hybrid DNN-HMMs with discriminative pre-training in comparison with other models, such as GMM-HMMs and MLP-HMMs. Another hybrid architecture combining DNN with an Extreme Learning Machine (ELM) was also successfully utilized for classifying emotion from utterances in [Han *et al.*, 2014].

Continuous prediction of dimensional emotion was investigated with a Deep Belief Network in [Schmidt *et al.*, 2012]. [Eyben *et al.*, 2012] dealt with multi-task recurrent neural network based speech emotion regression. The winning team of the last edition of the Audio-Visual Emotion recognition Challenge [Ringeval *et al.*, 2015b] employed bidirectional long short-term memory recurrent neural networks (DBLSTM-RNN) to perform unimodal emotion recognition and multimodal fusion [He *et al.*, 2015]. While these works use a similar learning framework as in our paper, none of them uses the discriminative objective based on correlation coefficients as introduced below.

3 Discriminative objectives for emotion regression

In the following, we will introduce two different objectives, one based on the CCC for each sequence, and one based on the overall CCC. We will denote by y_f^i the regression outputs for sequence i and target variable f (in case of neural networks, the sequences of activations of unit f of the output layer), while ${y_f^i}^*$ denotes the corresponding training targets (i.e., gold-standard). The standard sum of squared errors (SSE) training objective for a mini-batch $\mathcal{B}$ is given by

$$\sum_{i \in \mathcal{B}, f \in \mathcal{F}} \sum_t (y_{f,t}^i - {y_{f,t}^i}^*)^2, \tag{1}$$

where $\mathcal{F}$ is the set of target variables (e.g., arousal, valence, etc. in the case of emotion recognition) and t denotes the index of a time step at which the target variable is annotated.

While the above objective is discriminative, it is not discriminative on the sequence level. Let us thus introduce the proposed objective function based on the CCC per training sequence. The total cost function $\mathcal{O}$ is:

$$\mathcal{O} = - \sum_{i \in \mathcal{B}, f \in \mathcal{F}} \text{CCC}_f^i, \tag{2}$$

This objective will be denoted by ΣCCC below, as it is computed on all training sequences i available in the mini-batch

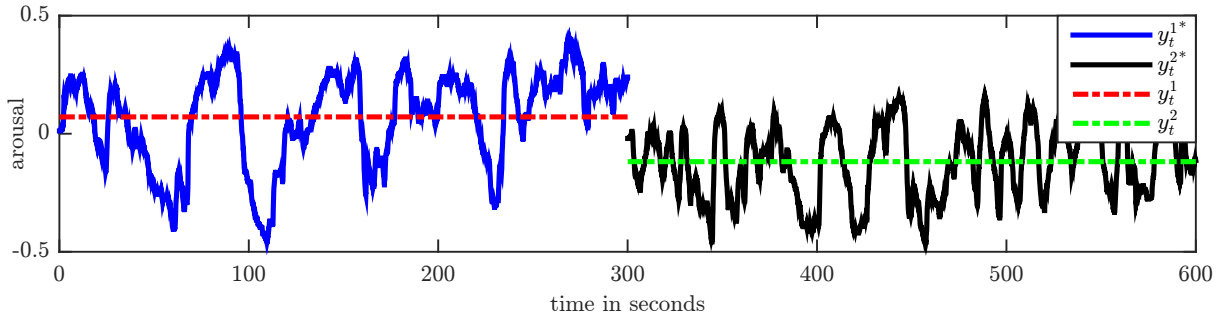

Figure 2: Illustration of the difference between CCC and $\sum$CCC on training targets $y_t^{i^*}$ for two sequences from the training partition of the RECOLA database (arousal) and predictions corresponding to the mean of each sequence: $y_t^i = \frac{1}{N_i}\sum_t^{N_i} y_t^{i^*}$. Because CCC is not a linear function, its computation between the predictions and the gold-standards provides different results whether we sum it over the two sequences ($\sum$CCC $= 0$), or compute it on the two concatenated sequences (CCC $= 0.37$).

$\mathcal{B}$. The CCC per sequence i and target f is defined in accordance with [Lin, 1989] as:

$$\mathrm{CCC}_f^i = \frac{2\mathrm{Cov}(y_f^i, {y_f^i}^*)}{\mathrm{Var}(y_f^i) + \mathrm{Var}({y_f^i}^*) + \Big(\mathrm{E}(y_f^i) - \mathrm{E}({y_f^i}^*)\Big)^2}, \tag{3}$$

where E, Var, and Cov denote sample mean, variance, and covariance, respectively.

Let us consider the mean squared error, $\mathrm{E}\{(y_f^i - {y_f^i}^*)^2\}$, which is equivalent to $\mathrm{Var}\{y_f^i - {y_f^i}^*\} + \mathrm{E}\{y_f^i - {y_f^i}^*\}^2 = \mathrm{Var}\{y_f^i\} + \mathrm{Var}\{{y_f^i}^*\} + \mathrm{E}\{y_f^i - {y_f^i}^*\}^2 - 2\mathrm{Cov}\{y_f^i, {y_f^i}^*\}$. Based on this observation, we can rewrite the CCC as:

$$\mathrm{CCC}_f^i = \frac{Q_f^i}{S_f^i + Q_f^i}, \tag{4}$$

defining N^i as the length of sequence i, the covariance-related quantity as:

$$Q_f^i := N^i \mathrm{Cov}\{y_f^i, {y_f^i}^*\} \tag{5}$$

$$= N^i \Big(\mathrm{E}\{y_f^i {y_f^i}^*\} - \mathrm{E}\{y_f^i\}\mathrm{E}\{{y_f^i}^*\}\Big) \tag{6}$$

$$= \sum_{t=1}^{N^i} y_{f,t}^i {y_{f,t}^i}^* - \mathrm{E}\{{y_f^i}^*\} \sum_{t=1}^{N^i} y_{f,t}^i, \tag{7}$$

and the sum of squared errors (SSE) related quantity as:

$$S_f^i := \frac{1}{2}\sum_{t=1}^{N^i} \Big(y_{f,t}^i - {y_{f,t}^i}^*\Big)^2. \tag{8}$$

An alternative objective to maximize (denoted simply by CCC below) is the 'total' CCC on the training set. This can be achieved by simply considering the entire training set as a single sequence i in (3). As shown in Fig. 2, the ΣCCC objective differs from the CCC objective in that it necessarily enforces accurate prediction of the target contour within each sequence, while the CCC objective could assign a good score to over-smoothed regression outputs that only predict the average label right. Conversely, if the target label has low variance within the sequences, the ΣCCC objective is hard to optimize and might emphasize on noise in the 'gold-standard', which is often given in emotion recognition. Thus, which objective is preferable certainly depends on the application.

Note that since the CCC on two partitions of the training set is not equivalent to the sum (or average) of the CCCs on these two partitions, it is not directly possible to optimize the total CCC on the training set, unless the mini-batch size comprises the total training set, which might be impractical. This is in contrast to SSE training and the ΣCCC maximization, where batch learning can be implemented by summing up the gradients from the mini-batches.

In the case of mini-batch learning with $|\mathcal{B}| = 1$ (one sequence per mini-batch), the optimization of ΣCCC and CCC is equivalent. However, in case of recurrent neural network training as considered here, $|\mathcal{B}| \gg 1$ is required for efficiency [Weninger *et al.*, 2015].

Further, we could also define an objective in analogy to (2) yet based on the Pearson's CC,

$$-\sum_{i\in\mathcal{B}, f\in\mathcal{F}} \mathrm{CC}_f^i = -\sum_{i\in\mathcal{B}, f\in\mathcal{F}} \frac{\mathrm{Cov}(y_f^i, {y_f^i}^*)}{\sigma_f^i \sigma_f^{i^*}}, \tag{9}$$

where σ denotes the standard deviations of outputs and targets in analogy to the above. However, since the Pearson CC is invariant w.r.t. the scale of the network outputs (in contrast to the CCC, cf. Fig. 1), this function has infinitely many minima, making its minimization hardly feasible.

4 Training algorithm

In this study, optimization of the discriminative objectives is performed by stochastic gradient descent. For the proposed CCC objective, we compute the gradient $\nabla_y \mathcal{O} = (\partial\mathcal{O}/\partial y_{f,t}^i)_{i,f,t}$. Using the quotient rule, the partial derivative for a sequence i, target f and time step t is computed as:

$$\frac{\partial \mathcal{O}}{\partial y_{f,t}^i} = -\frac{\frac{\partial Q_f^i}{\partial y_{f,t}^i} \cdot (S_f^i + Q_f^i) - Q_f^i \left(\frac{\partial S_f^i}{\partial y_{f,t}^i} + \frac{\partial Q_f^i}{\partial y_{f,t}^i}\right)}{\Big(S_f^i + Q_f^i\Big)^2} \tag{10}$$

with the partial derivatives

$$\frac{\partial Q_f^i}{\partial y_{f,t}^i} = {y_{f,t}^i}^* - \mathrm{E}\{{y_f^i}^*\}, \tag{11}$$

$$\frac{\partial S_f^i}{\partial y_{f,t}^i} = y_{f,t}^i - {y_{f,t}^i}^*. \tag{12}$$

With this, the desired derivative $\partial\mathcal{O}/\partial y_{f,t}^i$ is obtained as

$$\frac{({y_{f,t}^i}^* - \mathrm{E}\{{y_f^i}^*\})(S_f^i + Q_f^i) - Q_f^i(y_{f,t}^i - \mathrm{E}\{{y_f^i}^*\})}{\left(S_f^i + Q_f^i\right)^2} \tag{13}$$

$$= \frac{({y_{f,t}^i}^* - \mathrm{E}\{{y_f^i}^*\})S_f^i - (y_{f,t}^i - {y_{f,t}^i}^*)Q_f^i}{\left(S_f^i + Q_f^i\right)^2}. \tag{14}$$

Having obtained the output gradient as above, the gradient w.r.t. the weights, $\partial\mathcal{O}/\partial w$ is determined by backpropagation through time as usual.

Discriminative training is implemented on top of the open source, GPU-enabled neural network training software CURRENNT [Weninger *et al.*, 2015], which supports deep feed-forward and recurrent neural networks. The additional code which minimizes the proposed objectives will be made available upon publication of this manuscript.

During the *forward pass*, we compute and store the target means $\mathrm{E}\{{y_f^i}^*\}$, as well as S_f^i and Q_f^i in $|\mathcal{F}| \times |\mathcal{B}|$ matrices. The summations are computed using an outer loop over time steps, and using matrix addition for all i and f in parallel. For sequence lengths which are large in comparison to the number of target features and the batch size, this might still get inefficient. However, in our experiments, the speed of network training using the ΣCCC, CCC, or the standard SSE cost function was in the same ballpark. The *backward pass* is similar to the calculation of the SSE backward pass. The output gradient can be computed for all i, f, and t in parallel.

In case of optimizing CCC rather than ΣCCC, sequence boundaries need not be taken into account ($i = 1$). Consequently, the computation of the quantities S_f^1 and Q_f^1 for each f is much simpler and similar to the SSE calculation.

5 Experiments and Results

We present in this section the results of time-continuous dimensional emotion (arousal and valence) prediction tasks on two different corpora from different domains (speech and music). The objective of those experiments is to empirically demonstrate the benefits of using CCC as cost function for network training, in comparison to the traditional SSE, for emotion recognition from speech and music.

5.1 Emotions from music: MediaEval

Experiments on emotion recognition from music are done on the 'Emotion in Music Database' which was used in the MediaEval 2013 evaluation campaign [Soleymani *et al.*, 2013]. The task is to recognize time-varying emotion contours in the arousal and valence dimensions at a rate of 1 Hz from music signals. The data set includes excerpts of 45 seconds randomly selected (uniform distribution) from 744 songs taken from the online library Free Music Archive[1], and split between a development set (619 songs) and an evaluation set (125 songs). Ratings of emotion were performed on a crowdsourcing platform (MTurk) by a pool of 100 selected workers (57 male, 43 female, mean age is 32 years and standard deviation of 10 years) from different countries (72% from the USA, 18% from India and 10% from the rest of the world).

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

	Train	Dev	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)

Both features extraction and machine learning steps are based on the setup reported in [Weninger *et al.*, 2013]. The 6 373-dimensional ComParE set of generic affective features, and the Long Short-Term Memory (LSTM) [Gers *et al.*, 2000] architecture for deep recurrent neural networks (DRNNs) are used. LSTM networks have two hidden layers with 192 or 256 hidden units. The ComParE set consists of supra-segmental acoustic features, i.e., summarization of frame-level features over segments of constant length. In the present study, supra-segmental features are extracted from non-overlapping segments of 1 second length, in accordance with the time resolution of the emotion contours.

The training parameters are preserved from [Weninger *et al.*, 2013]. Input noise with $\sigma = 0.6$ is added to help generalization, and an early stopping strategy is used to alleviate overfitting. Stochastic gradient descent with a batch size of 25 sequences is used in all experiments. The learning rate η is determined in a preliminary cross-validation experiment for each objective function. Note that the objective functions are on different scales and hence the optimal step size varies between various objectives.

Both the sum of CCC and total CCC objectives are investigated. As baseline, standard SSE training is used. In accordance with the MediaEval challenge, the evaluation metrics comprise the overall Pearson's correlation coefficient (CC)[2] as well as the average Kendall's rank correlation coefficient per sequence ($\mathrm{E}\{\tau\}$), which is related to our ΣCCC objective function but not differentiable. Furthermore, we report the average CCC ($\mathrm{E}\{(\mathrm{CCC})\}$) per sequence, which directly corresponds to the ΣCCC objective.

[1] `http://www.freemusicarchive.org`

[2] Note that MediaEval uses the determination coefficient, which is the square of the CC, but we report CC as it is in the same order of magnitude as the CCC, which is the focus of our evaluation.

Table 2: Emotion recognition performance on the MediaEval 2013 test set (music domain). The best achieved Challenge metric (E$\{\tau\}$) is highlighted. Obj. denotes the objective function in network training and η the learning rate, determined in cross-validation.

			Arousal					Valence				
Layers	Obj.	η	CC	CCC	E{CCC}	E$\{\tau\}$	MLE	CC	CCC	E{CCC}	E$\{\tau\}$	MLE
192-192	SSE	10^{-5}	.795	.778	.148	.221	.136	.637	.632	.118	.189	.149
256-256	SSE	10^{-5}	.732	.724	.119	.174	.152	.623	.609	.109	.151	.142
192-192	CCC	10^{-2}	.792	.790	.149	.224	.140	.653	.648	.119	**.199**	.156
256-256	CCC	10^{-2}	.764	.761	.128	.161	.149	.648	.646	.130	.191	.154
192-192	ΣCCC	10^{-4}	.723	.719	.166	**.251**	.158	.547	.546	.136	.198	.168
256-256	ΣCCC	10^{-4}	.720	.717	.153	**.211**	.158	.587	.582	.130	**.198**	.158

5.2 Emotions from speech: RECOLA

Time-continuous prediction of emotional dimensions (arousal and valence) has also been investigated on speech data by using the RECOLA database [Ringeval *et al.*, 2013][3]; the full dataset was used for the purpose of this study, which corresponds to speech recordings from 46 French-speaking participants (27 female, 19 male, mean age is 22 years and standard deviation of 3 years) with 5 minutes for each. Ratings of emotion were obtained by 6 French-speaking research assistants (3 female, 3 female) using the ANNEMO annotation toolkit. Traces were then interpolated at a 40 ms frame rate and averaged as a gold-standard, using the same procedure as described in [Ringeval *et al.*, 2015a]. The dataset was split equally in three partitions – train (16 subjects), development (15 subjects) and test (15 subjects) – by stratifying (i.e., balancing) the gender, the age and the nationality of the speakers. Details on those partitions are provided in Table 1. The same procedure as the one used in the latest edition of the Audio-Visual Emotion Recognition Challenge (AV$^+$EC 2015) [Ringeval *et al.*, 2015b] has been used to extract acoustic features from the speech recordings: the extended Geneva minimalistic acoustic feature set (eGeMAPS – 102 features) [Eyben *et al.*, 2015] has been applied at a rate of 40 ms (to match the sampling frequency of the gold-standard) using overlapping windows of 3 seconds length.

For the prediction task, we used LSTM-DRNNs with three hidden layers with 128 units each. Input noise with $\sigma = 0.1$ is added and early stopping is also used to prevent overfitting. The networks were trained with stochastic gradient descent on a batch size of 5 sequences with a fixed momentum of 0.9, at different values of learning rate $\eta = \{10^{-2}, 10^{-3}, \ldots, 10^{-7}\}$. An optimal learning rate η was chosen based on the CCC on the development set for each emotional dimension and objective function. The CCC metric was computed on the gold-standard and prediction values concatenated over all recordings, in accordance with the AV$^+$EC challenge. In addition, we also report the average CCC (E{CCC}) per sequence in analogy to the experiments on music.

For all the networks (regardless of the training objective), a chain of post-processing was applied to the predictions obtained on the development set: (i) median filtering (with size of window ranging from 0.4 second to 20 seconds) [Ringeval *et al.*, 2015b], (ii) centring (by computing the bias between gold-standard and prediction) [Kächele *et al.*, 2015], (iii) scaling (using the ratio of standard-deviation of gold standard and prediction as scaling factor) and (iv) time-shifting (by shifting the prediction forward in time with values ranging from 0.04 second to 10 seconds), to compensate for delays in the ratings [Mariooryad and Busso, 2015]. Any of these post-processing steps was kept when an improvement was observed on the CCC of the development partition, and applied then with the same configuration on the test partition.

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

5.3 Results

Table 2 shows the results on the MediaEval 2013 test set (music). We can observe that the evaluation metrics exactly reflect the choice of the objective function: SSE training works best for minimizing the MLE, while CCC based training yields the best CCC on the test set.

The official Challenge evaluation metric, E$\{\tau\}$, is significantly (according to a z-test, $\alpha = .05$) improved by using the ΣCCC objective function (.221 $\rightarrow$ .251) for arousal but only slightly (.189 $\rightarrow$ .199) for valence. Generally, it is observed that the larger network with 256 hidden units performs worse on the test set, which can be attributed to the relatively small data set which causes over-fitting. The discrepancy between E{CCC} and CCC on this data set is astonishing; we found that for some test sequences, the variance in the annotated emotion contours is very low, which makes it hard to achieve good CC on these. One may further notice that the best performance measured as CC on valence is obtained with the CCC objective. The improvement over the SSE objective is significant (.637 $\rightarrow$ .653). Regarding the optimization of the network, results show that each objective function requires a specific learning rate to perform best.

Next, in Table 3 we report the metrics on the RECOLA database (speech). Here, we observe a significant improvement in the CC, CCC and E{CCC} metrics by using the ΣCCC objective function, particularly on the test set, where SSE training does not deliver useful results in the arousal dimension: CCC = .097 with SSE training and .350 with ΣCCC training. Since this difference is less pronounced on the development set, for which the network is tuned, we have some evidence that the ΣCCC objective function leads to better

Table 3: Emotion recognition performance on the RECOLA development and test partitions (speech domain). The best achieved Challenge metric (CCC) is highlighted. Obj. denotes the objective function in network training and η the learning rate, determined on the development results.

		Arousal					Valence				
Partition	Obj.	η	RMSE	CC	CCC	E{CCC}	η	RMSE	CC	CCC	E{CCC}
DEV	SSE	10^{-4}	.117	.412	.397	.227	10^{-4}	.105	.210	.201	.066
TEST	SSE	10^{-4}	.128	.109	.097	.161	10^{-4}	.108	.133	.131	.052
DEV	CCC	10^{-3}	.193	.373	.373	.294	10^{-2}	.133	.179	.179	.112
TEST	CCC	10^{-3}	.193	.257	.254	.212	10^{-2}	.130	.155	.155	.080
DEV	ΣCCC	10^{-5}	.217	.412	**.412**	.313	10^{-2}	.188	.249	**.242**	.150
TEST	ΣCCC	10^{-5}	.200	.351	**.350**	.268	10^{-2}	.192	.227	**.199**	.139

generalization. In fact, when training using the SSE criterion, we observed a bias of the network towards predicting the mean annotation on the training set, which leads to good RMSE but low correlation; conversely, the RMSE is significantly increased by using the CCC-based criteria. This result can also be observed on the CC evaluation metric, where a significant improvement over the SSE objective function is obtained when using $\sum$CCC for both arousal and valence. One may also note that the same results hold for the optimization of the network: each objective function requires a specific learning rate in order to provide the best performance.

6 Conclusions

We have demonstrated that the SSE objective in neural network regression can be effectively replaced by a criterion derived from the CCC, which has a significant impact on performance in continuous dimensional emotion recognition of arousal and valence from speech and music. Indeed, the CCC is an elegant solution to the issue of scaling and shifting time-continuous predictions, as it is sensitive to both of these variations and thus alleviates the problem of local minima in neural network training. Still, the observed increase in MSE-related criteria indicates that further investigations need to be performed in order to find an appropriate trade-off between MSE- and CC-like criteria.

Furthermore, note that the proposed approach based on CCC optimization can be applied to any sequence regression task where the correlation between the regression outputs and the ground truth should be maximized. There are no assumptions made on the underlying problem, other than that there be one or more continuous-valued target labels and that the regression model can be effectively trained by a first-order method such as stochastic gradient descent. Thus, we will verify its efficiency on other recognition tasks involving time-continuous measurements.

Finally, there are many more areas in emotion recognition and related fields where usage of the CCC can be explored, including the definition of the 'gold-standard' by computing the agreement of the raters, estimating an annotation delay, or the selection of features by using CCC instead of CC.

Acknowledgments

The research leading to these results has received funding from the European Commission's Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), and the European Union's Horizon 2020 Programme through the Innovative Action No. 644632 (MixedEmotions), No. 645094 (SEWA) and the Research Innovative Action No. 645378 (ARIA-VALUSPA).

References

[Coutinho and Cangelosi, 2010] Eduardo Coutinho and Angelo Cangelosi. A neural network model for the prediction of musical emotions. In S. Nefti-Meziani and J.G. Grey, editors, *Advances in Cognitive Systems*, pages 331–368. IET Publisher, London, UK, 2010.

[Eyben *et al.*, 2012] Florian Eyben, Martin Wöllmer, and Björn Schuller. A Multi-Task Approach to Continuous Five-Dimensional Affect Sensing in Natural Speech. *ACM Transactions on Interactive Intelligent Systems, Special Issue on Affective Interaction in Natural Environments*, 2(1), March 2012. 29 pages.

[Eyben *et al.*, 2015] Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 2015. in press.

[Gers *et al.*, 2000] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000.

[Han *et al.*, 2014] Kun Han, Dong Yu, and Ivan Tashev. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. In *Proc. of INTERSPEECH*, pages 223–227, Singapore, September 2014. ISCA.

[He *et al.*, 2015] Lang He, Dongmei Jiang, Le Yang, Ercheng Pei, Peng Wu, and Hichem Sahli. Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. In *Proc. of AVEC*, pages 73–80, Brisbane, Australia, October 2015. ACM.

[Huang *et al.*, 2015] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint Opti-

mization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(12):2136–2147, December 2015.

[Jenke *et al.*, 2013] Robert Jenke, Angelika Peer, and Martin Buss. A Comparison of Evaluation Measures for Emotion Recognition in Dimensional Space. In *Proc. of ACII*, pages 822–826, Geneva, Switzerland, September 2013. IEEE.

[Kächele *et al.*, 2015] Markus Kächele, Patrick Thiam, Günther Palm, Friedhelm Schwenker, and Martin Schels. Ensemble methods for continuous affect recognition: Multimodality, temporality, and challenges. In *Proc. of AVEC (held in conjunction with ACM MM)*, pages 9–16, Brisbane, Australia, October 2015. ACM.

[Kingsbury *et al.*, 2012] Brian Kingsbury, Tara N. Sainath, and Hagen Soltau. Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization. In *Proc. of ICASSP*, Kyoto, Japan, 2012.

[Li *et al.*, 2013] Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, I. Gonzalez, E. Valentin, and H. Sahli. Hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition. In *Proc. of ACII*, pages 312–317, Geneva, Switzerland, September 2013. IEEE.

[Lin, 1989] Lawrence I. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, March 1989.

[Mariooryad and Busso, 2015] Soroosh Mariooryad and Carlos Busso. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, 6(2):97–108, April-June 2015.

[Metallinou *et al.*, 2013] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2):137–152, February 2013.

[Ringeval *et al.*, 2013] Fabien Ringeval, Andreas Sonderegger, Jürgen Sauer, and Denis Lalanne. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proc. of EmoSPACE (held in conjunction with ACM FG)*, Shanghai, China, April 2013. 8 pages.

[Ringeval *et al.*, 2015a] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data. *Pattern Recognition Letters*, 66:22–30, November 2015.

[Ringeval *et al.*, 2015b] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. In *Proc. of AVEC (held in conjunction with ACM MM)*, pages 3–8, Brisbane, Australia, October 2015.

[Russell, 1980] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[Schmidt and Kim, 2011] Erik M. Schmidt and Youngmoo E. Kim. Modeling musical emotion dynamics with conditional random fields. In *Proc. of ISMIR*, pages 777–782, Miami, FL, USA, 2011.

[Schmidt *et al.*, 2012] Erik M. Schmidt, Jeffrey Scott, and Youngmoo E. Kim. Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion. In *Proc. of ISMIR*, pages 325–330, Miami, FL, USA, 2012.

[Schuller *et al.*, 2012] Björn Schuller, Michel Valstar, Florian Eyben, Roddy Cowie, and Maja Pantic. AVEC 2012 – The Continuous Audio/Visual Emotion Challenge. In Louis-Philippe Morency, Dan Bohus, Hamid K. Aghajan, Justine Cassell, Anton Nijholt, and Julien Epps, editors, *Proc. of ICMI*, pages 449–456, Santa Monica, CA, October 2012. ACM.

[Soleymani *et al.*, 2013] Mohammad Soleymani, Michael N. Caro, Erik M. Schmidt, Chen-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proc. of CrowdMM (held in conjunction with ACM MM)*, Barcelona, Spain, 2013. ACM.

[Veselý *et al.*, 2013] Karek Veselý, Arnab Ghoshal, Lukáš Burget, and Daniel Povey. Sequence-discriminative training of deep neural networks. In *Proc. of INTERSPEECH*, pages 2345–2349, Lyon, France, 2013. ISCA.

[Wang *et al.*, 2015] Fengna Wang, Hichem Sahli, Junbin Gao, Dongmei Jiang, and Werner Verhelst. Relevance units machine based dimensional and continuous speech emotion prediction. *Multimedia Tools and Applications*, 74(22):9983–10000, November 2015.

[Weninger *et al.*, 2013] Felix Weninger, Florian Eyben, and Björn Schuller. The TUM approach to the MediaEval music emotion task using generic affective audio features. In Martha Larson, Xavier Anguera, Timo Reuter, Gareth J.F. Jones, Bogdan Ionescu, Markus Schedl, Tomas Piatrik, Claudia Hauff, and Mohammad Soleymani, editors, *Proc. of MediaEval*, Barcelona, Spain, October 2013. CEUR.

[Weninger *et al.*, 2014] Felix Weninger, John R. Hershey, Jonathan Le Roux, and Björn Schuller. Discriminatively trained recurrent neural networks for single-channel speech separation. In *Proc. of GlobalSIP*, pages 740–744, Atlanta, GA, USA, 2014. IEEE.

[Weninger *et al.*, 2015] Felix Weninger, Johannes Bergmann, and Björn Schuller. Introducing CURRENNT – the Munich open-source CUDA RecurREnt Neural Network Toolkit. *Journal of Machine Learning Research*, 16:547–551, 2015.