Reconstruction-error-based learning for continuous emotion recognition in speech

Jing Han, Zixing Zhang, Fabien Ringeval, Björn Schuller

RECONSTRUCTION-ERROR-BASED LEARNING FOR CONTINUOUS EMOTION RECOGNITION IN SPEECH

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

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

ABSTRACT

To advance the performance of continuous emotion recognition from speech, we introduce a reconstruction-error-based (RE-based) learning framework with memory-enhanced Recurrent Neural Networks (RNN). In the framework, two successive RNN models are adopted, where the first model is used as an autoencoder for reconstructing the original features, and the second is employed to perform emotion prediction. The RE of the original features is used as a complementary descriptor, which is merged with the original features and fed to the second model. The assumption of this framework is that the system has the ability to learn its 'drawback' which is expressed by the RE. Experimental results on the RECOLA database show that the proposed framework significantly outperforms the baseline systems without any RE information in terms of Concordance Correlation Coefficient (.729 vs .710 for arousal, .360 vs .237 for valence), and also significantly overcomes other state-of-the-art methods.

Index Terms— Continuous emotion recognition, reconstruction error, bidirectional long short-term memory

1. INTRODUCTION

Automatic continuous emotion recognition in speech is one of the most active research areas in the affective computing community in recent years [1, 2]. Many efforts have been reported in order to advance the system performance and further facilitate its applications in the real world [3–6]. Among these efforts, a memory-enhanced Recurrent Neural Network (RNN), namely *Long Short-Term Memory* (LSTM) RNN, has attracted considerable attention to address the raised task in the past few years [7–12]. The successful implementation of LSTM-RNN is mainly due to its powerful capability of learning long-range contextual information for sequential patterns [13, 14].

More recently, advanced research has begun to explore the benefits of LSTM in the context of continuous emotion recognition. In [6], Weninger et al. proposed a novel discriminative learning method, which exploits the Concordance Correlation Coefficient (CCC), rather than the traditional Root Mean Square Error (RMSE), as a differentiable objective function to train a LSTM-RNN model. In [15], Trigeorgis et al. utilised data-learnt high-level representations, instead of the traditional hand-crafted features like Mel-Frequency Cepstral Coefficients (MFCC), as the inputs of LSTM-RNN. The high-level representations are derived from the raw speech signals with the aid of Convolutional Neural Networks (CNN), which are jointly combined with the LSTM-based recognition model to constitute an end-to-end framework.

In this paper, we propose another novel framework based on the *Reconstruction Error* (RE) for continuous emotion recognition in speech. The underlying assumption is that, the LSTM-RNNs could learn their 'drawbacks' if we provide them a way to know when the inputs are more challenging to process. Specifically, we utilise the RE from one LSTM-RNN model, used as an Auto Encoder (AE), to express the drawbacks of another LSTM-RNN model used for emotion regression. The RE is thus used as a complementary descriptor of the original features, that describes the ability of the LSTM-RNN model to encode the emotion related information.

The idea of the proposed method is partially inspired by the promising attention modelling in the field of natural language processing [16], where the decoder pays different attention to the regions of the input based on their relevant degree with the output. As to our method, the continuous variation of the RE provides different levels of disturbance when training a model with time-continuous features. In this contribution, we demonstrate by empirical analysis that this approach can help heighten the attention of a model on the error-sensitivity regions, so as to ameliorate the model performance for emotion recognition from speech.

2. RELATED WORK

RE has been originally used as an objective function of AE for extracting high-level representations. Recently, however, a trend in the machine learning community has emerged towards exploiting directly the RE as descriptors for other tasks, e. g. , it was considered as a novelty measure for novelty detection [17]. An AE is trained on normal samples beforehand to serve as a novel event detector. When a new sample is fed into the AE, its RE is compared with a predefined threshold to decide whether the current sample is abnormal. When an unknown sample passes through the AEs that are trained class-specifically, the corresponding RE indicates the likelihood to belong to a class.However, to the best of our knowledge, none of the works takes the RE into consideration for augmenting the learning capability of another similar model for recognition.

This work is also relevant to the tandem architectures in the speech recognition domain, which use the state or phoneme posterior probabilities generated by a neural network as features observed by a Hidden Markov Model (HMM). Thanks to the development of neural networks, the topologies have been shifted from the early Multilayer Perceptron (MLP) [18] to the recent LSTM-RNN [19]. For emotion recognition, and to the best of our knowledge, only the tandem structure of LSTM-DBN has been exploited in [20] so far.

In this work, the advantages of disparate models, i. e., the context-sensitivity capabilities of LSTM, and the generalisation competence of HMM or DBN, have been explored.

Nevertheless, these architectures only take the prediction or the prediction probability as another dimensional representation, which are merely designed for classification tasks rather than regression problems. In this paper, we leverage for the first time the RE information of an LSTM-based AE to perform a time-continuous regression task (emotion) from speech.

3. RECONSTRUCTION-ERROR-BASED LEARNING

3.1. Overview

The framework of our proposed RE-based learning is depicted in Fig. 1, which consists of two stages: (i) the *extraction* stage, and (ii) the *exploitation* stage. In the first stage, the RE information is collected by applying a first model to rebuild each frame of the input as an AE. Given the inputs $\boldsymbol{x}_t$ of the model at time t, and its output counterparts $\hat{\boldsymbol{x}}_t$, the RE ϵ_t can be calculated by $\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\|$. In the second stage, the RE information combined with original features, i. e., $[\boldsymbol{x}_t, \epsilon_t]$, are used as new inputs to build a regression model that performs emotion prediction.

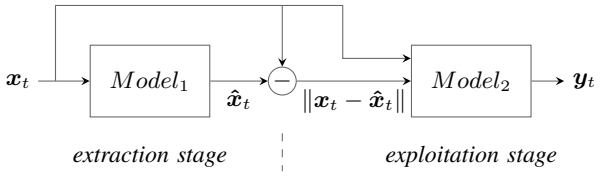

Fig. 1. Framework of reconstruction error based learning

One key question of this framework is how to guarantee that the RE information, that is derived from the first model, can be successfully exploited by the second model. To this end, we assume that the two models have a similar structure such that the RE extracted with the first model can well reflect the drawbacks of the second model. In this paper, we employ Bidirectional LSTM-RNN (BLSTM-RNN) as the basic model because of its great success in continuous recognition of emotion, cf. Section 1.

In general, the BLSTM-RNN structure is composed of one input layer, one or multiple hidden layers, and one output layer [13]. The bidirectional hidden layers separately process the input sequences in a forward and a backward order and connect them to the same output layer. Compared with conventional RNNs, it adopts LSTM blocks to replace the neurons in the hidden layers. Each block consists of a self-connected memory cell and three gate units, namely input, output, and forget gate. These three gates allow the network to learn when to write, read, or reset the value in the memory cell. Such a structure grants BLSTM-RNN to learn past and future context in both short and long range. For a more in-depth explanation of BLSTM-RNNs the reader is referred to [13].

3.2. Extraction Stage

To extract the RE information, the BLSTM-RNN model is trained in a completely unsupervised way. That is, the inputs and the targets are exactly the same. If the model would be sufficiently powerful, one would expect that all inputs should be recovered. However, many empirical experiments have shown that the results are far from this expectation [10], which somewhat implies that BLSTM-RNN has its own drawbacks, just like any other machine learning technique.

Specifically, given a time sequence as input, the BLSTM-RNN is trained to minimise the cost function as

$$\mathcal{J}(\boldsymbol{\theta}) = \sum_{t=1}^{T} (\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t)^2, \tag{1}$$

where $\boldsymbol{x}_t$ is a sample at time t from an input sequence lasting a period of time T, and $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$ denotes the reconstructed sample of the corresponding input $\boldsymbol{x}_t$.

Once the first BLSTM-RNN model is trained, the RE ϵ_t can be obtained by computing the sum of the Euclidean distance between the input $\boldsymbol{x}_t$ and its corresponding reconstruction $\hat{\boldsymbol{x}}_t$ over all L dimensions, as expressed in Eq. (2). The result, ϵ_t, is therefore a scalar.

$$\epsilon_t = \sum_{l=1}^{L} |\hat{x}_{t,l} - x_{t,l}|, t \in T. \tag{2}$$

3.3. Exploitation Stage

In this stage, it is expected that the features with their respective RE values will attract different attention from the regression model. After the RE information ϵ_t is generated, it is simply concatenated with the original features $\boldsymbol{x}_t$ to form a longer feature vector $[\boldsymbol{x}_t, \epsilon_t]$, which thus incorporates information regarding reconstruction issues obtained with the first model. The conventional machine learning paradigms can be then applied sequentially.

It is worth noting that the features $\boldsymbol{x}_t$ used for training $Model_1$ in the extraction stage could be either frame-based Low-Level Descriptors (LLD), or the segment-based statistical features – after applying functionals over the LLD –, which may result in a difference in the final feature dimensions used for the exploitation stage. More details on these settings are explained in Section 4.2.

4. EXPERIMENTS AND RESULTS

In this section, we present a set of experiments on a time- and value-continuous dimensional emotion (arousal and valence) prediction task. The objective of the experiments is to empirically demonstrate the benefits of adding RE information to the conventional features for the proposed task.

4.1. Selected Database and Features

We evaluate our RE-based learning method on the RECOLA [21] database, which has been adopted for the AudioVisual Emotion recognition Challenges (AVEC) in 2015 [1] and 2016 [2]. This database contains spontaneous and natural interactions between remote participants that solved a task in dyads. The whole dataset contains recordings from 46 different subjects. The dataset is further divided into three disjoint partitions while balancing the gender, age, and mother tongue of the participants (cf. Table 1). It is worth mentioning that, we used exactly the same partitions as in [6, 15].

To annotate the corpus, time- and value-continuous dimensional affect ratings in terms of arousal and valence were performed by six French-speaking raters (three females) for the first five minutes of all recorded sequences. The obtained labels were then resampled at a constant frame rate of $40\,ms$, and averaged over all raters by considering inter-evaluator agreement, to provide a 'gold standard' [21]. Measures of inter-rater agreement show the reliability of the annotation data and the used post-processing techniques [1, 2].

Table 1. Three partitions of the RECOLA database.

#	train	development	test
female/male	10/6	9/6	8/7
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)

To extract acoustic features from the speech recordings, we used our open-source openSMILE toolkit [22] to extract 13 LLDs, i. e., MFCC 0–12 and logarithmic energy, with a frame window size of $25\,ms$ and a step of $10\,ms$. The arithmetic mean and the coefficient of variance were then computed over the sequential LLDs at a rate of 40 ms – to match the granularity of the annotation – using overlapping windows of 8 s length, resulting in 26 statistical features per analysis window. The total numbers of segments in the train, development, and test partitions are $120\,000$, $112\,500$, and $112\,500$, respectively.

4.2. Implementation and Evaluation

When training the first BLSTM-RNN model as an AE, there are two methods for choosing the training data as mentioned in Section 3.3 – either by the 13 LLDs (*lld-based* strategy), or by the 26 statistical features (*functional-based* strategy). This choice defines then the number of nodes of the input and output layers of the model, i. e., either 13 or 26, respectively. In both cases, only one dimensional RE ϵ_t is produced per input by Eq. (2), giving rise to 14 and 27 dimensional features, respectively. However, for the LLDs, functionals have to be applied before feeding them into the second BLSTM-RNN model. In doing this, there are 28 statistical features obtained in total for the lld-based strategy, since the means and variations are further calculated over the 13 LLDs plus the generated RE ϵ_t. Therefore, when training the second model to perform emotion prediction, the number of nodes of its input layer is either 28 for the lld-based strategy, or 27 for the functional-based strategy, respectively.

We adopted two hidden layers for both BLSTM-RNN models. In our experiments, the first model consists of 30 nodes per hidden layer with a learning rate of 10^{-6}, whereas 26 nodes are used for the second model, with a learning rate of 10^{-5}. Zero mean Gaussian noise with standard deviation 0.2 was added to the input activations in the training phase of both models to improve generalisation. The parameters of each model were optimised on the validation set with an early stopping strategy. Note that, an online standardisation was carried out on the features for both validation and test partitions, i. e., the means and variances of the features were calculated on the training partition and used on the two other partitions for standardisation. Additionally, annotation delay compensation was performed to compensate for the temporal delay between the observable cues, as seen in the recordings, and the corresponding emotion reported by the annotators [23]. As in [2, 24], we identified this delay to be four seconds which was duly compensated, by shifting the gold standard back in time with respect to the features for both arousal and valence.

To evaluate the performance of our methods, we used the CCC, which is a standard evaluation metric for time- and value-continuous predictions of emotion [1]; it measures here the agreement between the gold standard and the prediction provided by the second BLSTM-RNN model. Given two time sequences x and y, their CCC is defined as follows:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{3}$$

where ρ is the *Pearson's Correlation Coefficient* (PCC) between two time series (e.g., prediction and gold-standard); μ_x and μ_y are the means of each time series; and σ_x^2 and σ_y^2 are the corresponding variances. In contrast to the PCC, CCC takes not only the linear correlation, but also the bias between the two temporal series, i.e., $(\mu_x - \mu_y)^2$, into account. Hence, the value of CCC is within the range of $[-1, 1]$, where ± 1 represents perfect concordance and discordance while 0 means no correlation.

Table 2. Performance comparison (CCC) between the baseline and the proposed reconstruction-error-based learning methods (*lld-* or *functional*-based) on the dev(elopment) and *test* partitions for both arousal and valence regression. The symbol of $*$ indicates the significance of the performance improvement over the baseline.

	arousal		valence	
CCC	dev	test	dev	test
baseline	.712	.703	.333	.238
lld-based	.762*	.723*	.360*	.263*
functional-based	.710	.714*	.368*	.298*

To further assess the significance level of performance improvement, a statistical evaluation was carried out over all predictions obtained with the proposed RE-based learning and with a baseline method (i.e., original features without RE), by means of the Fisher's r-to-z transformation [25]. Unless stated otherwise, a p value lower than .05 indicates statistical significance.

4.3. Results and Discussion

The performance of the proposed RE-based learning with two RE extraction strategies (i.e., lld-based or functional-based) on the development and the test sets, for both arousal and valence regression tasks, is shown in Table 2. One can notice that in almost all cases, the RE-based learning methods significantly outperform the baseline systems that do not include the RE information. Even though a slightly decrease of performance can be observed on the development set for arousal and the functional based approach, the performance on the test set still benefits from the use of RE as additional information.

These results are highly encouraging since we only added a summary of the ability of the LSTM-RNN model to encode emotion related features into the original feature set, which yet contributes to a significant improvement of performance. In addition, we tentatively combined the RE features from another unmatched model rather than BLSTM-RNN (e. g., Support Vector Regression), no similar performance improvement was observed from the second BLSTM-RNN model. This might imply that the RE extracted from a similar structure better represent its drawbacks and is more helpful for the exploitation.

Further, following the AVEC's post-processing procedure of predictions [1, 2], that was successfully replicated in other studies [6, 15], we applied the same chain of post-processing on the obtained predictions; smoothing, centring, scaling, and time-shifting. All the modification parameters are optimised on the development set.

The resulting performance for both RE-based learning methods and the baseline system is presented in Table 3. As expected, all

Table 3. Performance comparison (CCC) between the proposed reconstruction-error-based learning methods (*lld-* or *functional*-based), the baseline system, and two other approaches (LSTM with CCC as cost function, and end-to-end learning), on the dev(elopment) and *test* partitions for both arousal and valence regression, after post-processing the predictions. The symbol of $*$ indicates the significance of the performance improvement over the baseline.

	arousal		valence	
CCC	dev	test	dev	test
baseline	.776	.710	.333	.237
LSTM [6]	.412	.350	.242	.199
End-to-End [15]	.741	.686	.325	.261
lld-based	**.785***	**.729***	.364*	.309*
functional-based	.754	.720*	**.378***	**.360***

of the post-processed results performed better than the results without post-processing of the predictions. In this scenario, one can still observe that the RE-based learning methods persistently outperform the baseline by a significant margin. Particularly, the best results on the test set are achieved at .729 of CCC for arousal with the lld-based strategy, and at .360 of CCC for valence for the functional-based strategy. Additionally, the RE-based learning methods remarkably outperforms the state-of-the-art methods, which are investigated in [6, 15] and outlined in Section 1.

To further investigate whether the proposed features contribute to the observed performance improvement, we calculate the PCC between the RE-based features and the performance improvement, and the PCC between the RE-based features and the gold standard. Specifically, the performance improvement is defined as $\delta = |p_b - gs| - |p_\epsilon - gs|$, given the golden standard (gs) and the predictions of the input by using the baseline method (p_b) or the RE-based methods (p_ϵ). Furthermore, for the lld-based strategy, the RE-based feature is calculated by $\epsilon = 0.5 * std(\epsilon_t) + 0.5 * var(\epsilon_t)$; while for the functional-based strategy, $\epsilon = \epsilon_t$.

The corresponding values of the PCC are presented in Table 4. From the table, one can see that the RE features have a relationship with the performance improvement, particularly for valence regression. This somewhat implies that the second BLSTM-RNN model pays more attention to the input regions that have high RE values, and yields better performance in these regions. This also means that the model holds the capability of learning from its drawbacks, as we hypothesised at the beginning of the paper. Furthermore, the PCC between the RE features and the performance improvement is higher for valence than the values for arousal, which confirms the observation in Table 2 and 3 that more improvement was obtained for valence regression than for arousal when integrating the RE-based features. In addition, one can also notice that the RE-based features have some correlation with the gold standard for both arousal and valence regressions. This indicates that the RE-based features take some pattern information which might be complementary for the original features.

To highlight the performance of the RE-based learning methods, Fig. 2 illustrates the automatic predictions of arousal and valence obtained with the best settings among the two proposed strategies (lld- or functional-based) for a single test subject. In general, the predictions generated by the proposed method are closer to the gold standard, which consequently contributes to better results in terms of CCC.

Table 4. Linear correlations (PCC, ρ) between the reconstruction error (ϵ) and the *prediction improvement* (δ), or between the reconstruction error (ϵ) and the *golden standard* (gs), on the *test* partition for both arousal and valence regression.

	arousal		valence	
PCC	$\rho(\epsilon, \delta)$	$\rho(\epsilon, gs)$	$\rho(\epsilon, \delta)$	$\rho(\epsilon, gs)$
lld-based	.086	.372	.561	.208
functional-based	.019	.417	.467	.200

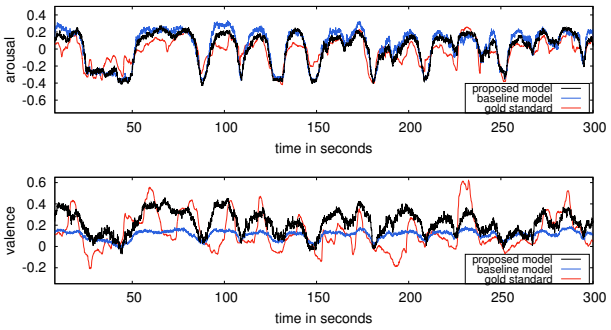

Fig. 2. Illustration of arousal and valence predictions obtained by the reconstruction-error-based learning method and the baseline method for the 4th subject from the test partition.

5. CONCLUSIONS

In this paper, we proposed a novel Reconstruction-Error-based (RE-based) learning framework for continuous emotion recognition in speech. It extracts the RE information from one learning model, then regards this information as additional features for another similar learning model for regression. Experiments were performed on the spontaneous emotional database – RECOLA – with BLSTM-RNN. Two variants were investigated, as RE information could be extracted from either the frame-based Low-Level Descriptors (LLD) or the segment-based statistical features after applying functionals on the LLD. When combining the RE information with the original features, the performance of the continuous emotion recognition systems can be significantly improved, and perform even better than the most recent approaches. Moreover, the high correlation between the RE and the performance improvement indicates that the RE information has a positive impact on the model. One may further note that, the proposed methods could also be applied to other regression problems, or even for classification tasks. For future work, we will investigate into more details the reasons that cause the BLSTM-RNN model to fail in the reconstruction, and will exploit other popular regression models, like Support Vector Regression.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

[1] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AVEC 2015: The 5th international audio/visual emotion challenge and workshop," in *Proc. ACM MM*, Brisbane, Australia, 2015, pp. 1335–1336.

[2] M. F. Valstar, J. Gratch, B. W. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016 - depression, mood, and emotion recognition workshop and challenge," in *Proc. the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Amsterdam, The Netherlands, 2016.

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

[4] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, Feb 2013.

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

[6] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio," in *Proc. IJCAI*, New York, NY, 2016, pp. 2196–2202.

[7] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies." in *Proc. INTERSPEECH*, Brisbane, Australia, 2008, pp. 597–600.

[8] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, Feb 2013.

[9] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proc. the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Brisbane, Australia, 2015, pp. 65–72.

[10] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks," in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 3593–3597.

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

[12] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Brisbane, Australia, 2015, pp. 73–80.

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

[14] A. Graves, *Supervised sequence labelling with recurrent neural networks*. Berlin/Heidelberg, Germany: Springer, 2012, vol. 385.

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

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. International Conference on Learning Representation (ICLR)*, San Diego, CA, 2015, 15 pages.

[17] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks," in *Proc. ICASSP*, Brisbane, Australia, 2015, pp. 1996–2000.

[18] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 1635–1638.

[19] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework," *Cognitive Computation*, vol. 2, no. 3, pp. 180–190, Apr 2010.

[20] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867–881, Oct 2010.

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

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

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

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

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