END-TO-END SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORKS

Panagiotis Tzirakis[1*], *Jiehao Zhang*[1*], *Björn W. Schuller*[1,2]

[1] Department of Computing, Imperial College London, London, UK
[2] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
panagiotis.tzirakis12@imperial.ac.uk

ABSTRACT

Affect recognition is an important component towards the better interaction between human and machines. Applications of emotion recognition in speech can be found in several areas such as human computer interaction and call centres. In recent years, Deep Neural Networks (DNN) have been used with great success in recognizing emotions. In this paper, we present a new model for continuous emotion recognition from speech. Our model, which was trained end-to-end, is comprised of a Convolutional Neural Network (CNN), which extracts features from the raw signal, and stacked on top of it a 2-layer Long Short-Term Memory (LSTM), so as to consider the contextual information in the data. Our model significantly outperforms, in terms of concordance correlation coefficient, the state-of-the-art methods for the RECOLA database.

Index Terms— speech emotion recognition, deep learning, end-to-end learning

1. INTRODUCTION

Recognising emotions is automatically and subconsciously performed by humans. It is a vital process for human-to-human communication, and thus, to achieve better human-machine interaction, emotions need to be considered. There are three major approaches for quantifying emotions, namely, categorical, continuous and appraisal-based. A popular continuous emotional model is the circumplex of affect [1], which models emotions using two independent dimensions, i.e. arousal (relaxed vs aroused) and valence (pleasant vs unpleasant).

However, emotion recognition is a challenging task as human emotions have fuzzy temporal boundaries. Difficulties arise into specifying the start or the completion of an emotion. In addition, emotions are expressed differently for each individual, and one utterance may contain more than one emotion.

Deep Neural Networks (DNNs) have emerged the recent years and had groundbreaking improvements in different areas of machine learning including the continuous affect recognition domain. Numerous new DNNs architectures have been proposed recently towards that direction such as Convolutional Neural Networks (CNNs), and Long-Short Term Memory (LSTM) networks.

A number of studies in the literature have focused on predicting emotion from speech using DNNs. Wllmer et al. [2] were one of the first to propose a DNN architecture for Affective computing which comprised of a three layer LSTM and was trained based on functionals of acoustic Low-Level Descriptors (LLDs). Stuhlsatz et al. [3] used Restricted Boltzmann Machines (RBM) to extract discriminative features from the raw signal and proposed a Generalized Discriminant Analysis (GerDA).

These studies use hand-crafted features to feed their DNN architectures. In this paper, we propose a new *end-to-end* convolution recurrent neural network architecture for continuous affect recognition. Despite most of the studies in the literature, the creation of our network architecture was inspired by the way conventional speech features like Mel-Frequency Cepstral Coefficients (MFCCs), are computed. Finally, our model surpasses the state-of-the-art studies for the RECOLA database.

The rest of the paper is structured as follows. Section 2 provides the most recent studies related to our work. Section 3 introduces our model and how it is related to conventional speech features. After the description of the dataset used in this study (Section 4), we present our results Section 5.

2. RELATED WORK

A number of studies have been proposed to model the raw waveform directly from a DNN. More specifically, Dielman et al. [4] trained a CNN on the raw audio signal and concluded that the network can find both frequency decompositions and phase invariant features. In another study, Sainath et al. [5, 6] proposed a a Convolutional, Long Short-Term Memory Deep Neural Network (CLDNN) model for a speech recognition task, that is able to reduce temporal and frequency variations. Dai et al. [7] proposed an end-to-end very deep neural network to extract features to learn acoustic models.

Several studies have also been proposed for recognising affect. For example, Schmitt et al. [8] used a bag-of-audio-

*The authors share joint first authorship.

words (BoAW) approach that was created from MFCCs and energy Low Level Descriptors (LLDs), as feature vector and a simple Support Vector Regression (SVR) to predict the arousal and valence dimensions. Other studies have used DNNs for this task. An example is the study of Trigeorgis et al. [9] that proposed an end-to-end model which comprised of a CNN architecture used to extract features before feeding a Bi-directional LSTM (BLSTM) to model the temporal dynamics in the data. In another study, Neumann et al. [10] propose an attentive convolutional neural network (ACNN) that combines CNNs with attention. The experimented with four different input feature sets, namely, 26 logMel filterbanks, 13 MFCCs, a prosody feature set, and (d) the extended Geneva minimalistic acoustic parameter set (eGeMAPS) [11].

3. MODEL DESIGN

The key operation of our model is convolution,

$$(f * h)(t) = \sum_{k=-T}^{T} f(t) \cdot h(t-k), \tag{1}$$

where $f(x)$ indicates a kernel function, which in our case operates on the raw signal $h(k)$.

To reduce the dimensionality of the signal, we utilise the max-pooling operation. To the best of our knowledge, none of the studies in the literature mention a methodology on how the pooling size should be selected based on the kernel size of the convolution layer. In most cases, these are selected experimentally.

In this work, we propose a simple methodology that is inspired on the calculation of conventional features. More specifically, the rate of overlap (R) between kernel size (K) and pooling size (P) is

$$R = \frac{K-1}{K+P-1}. \tag{2}$$

The overlapping ratio R is obviously less than 1 and its common value for creating hand-crafted features is considered most of the times around 0.5. If we use stride instead of max-pooling to reduce the dimensionality of the signal, we need to keep the rate of overlap around 0.5. However, we found that using stride provides worse performance than using max-pooling. When using max-pooling, we extract the most important information and discard the futile, while it takes all information into consideration when using strides. At this time, we need to keep R less than 0.5 since we do not want it to extract the same features for successive frames. To create our architecture we utilise this factor, so between convolution and max-pooling layers we consider $R < 0.5$ (i.e. $R \approx 0.4$) for all layers.

3.1. Proposed Model

Our proposed model, which is depicted in figure 1, is described below.

Input. After we preprocess the raw signal to have zero mean and unit variance, we segment it to 20 s sequences long and use them as input. At 16 kHz this corresponds to 320 000-dimensional input vector.

Temporal Convolution. We use 64 time impulse filters with kernel size of 8 to extract information from the raw signal.

Max-pooling. Based on the previous kernel size (8), we apply max-pooling with a size of 10 to decrease the frame rate of the signal and keep the most descriptive features.

Temporal Convolution. Going deeper we want to extract a larger number of high-level abstractions. For this purpose, we convolve in the time domain with a kernel size of 6 and with a channel size of 128.

Max-pooling. To keep the overlap rate below 1, we pool across the time domain with a size of 8.

Temporal Convolution. The last convolution layer provides us with an even higher level of abstractions. We keep the kernel size to 6 and increase the filter size to 256.

Max-pooling. Considering the kernel size of the previous convolution layer, we perform a max-pooling across time with a size of 8.

Recurrent Neural Network. We utilise 2-layers of LSTM to capture the contextual information in the data.

Due to the high number of parameters our model contains we use dropout regularisation after each pooling layer, with a probability of 0.5.

3.2. Objective function

Most of the studies in the literature use the Mean Squared Error (MSE) as a loss function to train their model. However, a new trend in the speech analysis domain has emerged to use a loss function based on the concordance correlation coefficient (CCC), which has been shown to provide better results [9, 12]. We utilise the same loss function ($\mathcal{L}_c$) based on the CCC (ρ_c).

$$\mathcal{L}_c = 1 - \rho_c = 1 - \frac{2\sigma_{xy}^2}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3}$$

$$= 1 - 2\sigma_{xy}^2 \psi^{-1}, \tag{4}$$

where $\psi = \sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2$ and $\mu_x = E(x)$,$\mu_y = E(y)$,$\sigma_x^2 = var(x)$,$\sigma_y^2 = var(y)$ and $\sigma_{xy}^2 = cov(x, y)$. The gradient of the loss to be propagated from the last layer with respect to the weights is

$$\frac{\partial \mathcal{L}_c}{\partial x} \propto 2\frac{\sigma_{xy}^2(x - \mu_y)}{\psi^2} + \frac{\mu_y - y}{\psi}. \tag{5}$$

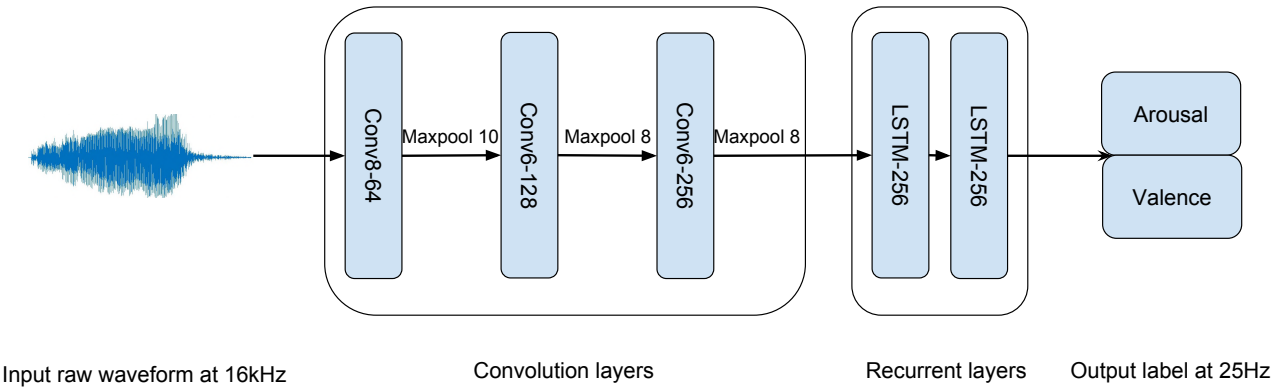

Fig. 1. The proposed convolutional recurrent neural network for speech emotion recognition. A CNN is used to extract features from the raw signal before feeding them to a 2-layer LSTM network for the final prediction.

4. DATASET

To test our methodology and architecture, we utilise the REmote COLlaborative and Affective (RECOLA) database introduced by Ringeval et al. [13]. A subset of the database was used in the Audio/Visual Emotion Challenge and Workshop (AVEC) 2015, and 2016 challenges [14, 15]. However, in this study, we utilise its full portion, which contains 46 different recordings divided into three different parts (train, devel, and test) while balancing the gender, age and mother tongue. Four modalities are contained in the database, namely, audio, video, electrocardiogram (ECG), and electro-dermal activity (EDA). The original labels of the RECOLA are re-sampled at constant frame rate of 40 ms. The data is then averaged over all raters by considering the inter-evaluator agreement, to provide a gold standard [16]. Each record is 300 s audio data with a sampling rate of 16 kHz. Table 1 contains more details for each portion of the dataset.

	Train	**Valid**	**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 $\mu(\sigma)$	22.3(3.4)	21.6(2.1)	21.2(2.0)

Table 1. The partitioning of the RECOLA dataset

5. EXPERIMENTS AND RESULTS

5.1. Experimental Setup

The optimisation method we used to train our model, throughout all experiments, is the RMSProp optimizer [17] with a fixed learning rate of 10^{-4}, a decay rate of 0.9, and momentum of 0.1. The mini-batch size utilised was 5 with a sequence length of 500 frames (20 s) when training and the model is tested on the entire records without segmentation. In addition, as mentioned earlier, regularisation was used to prevent overfitting, and in particular dropout [18] was used after the max-pooling layers with probability 0.5. The model is selected based on the highest CCC before post processing on the validation partition. Finally, a chain of post processing method is applied, namely, median filtering (size of window was between 0.04 s and 20 s) [14], centering (by finding the ground truth's and the prediction's bias) [19], scaling (with scaling factor the ration between the standard deviation of the ground truth and the prediction) [20] and time-shifting (forward in time with values between 0.04 s and 10 s) [21]. Any of these method is kept when we observe a better ρ_c on the validation set, and then applied to the test partition with the same configuration.

5.2. Results

We first compare ourselves with the study performed by Trigeoris et al. [9] as they also use DNNs in an end-to-end manner. We should note here that they consider a sequence length of 6 s as input to the network. However, we increase the length of the input sequence (i.e., 20 s) so that we can make the model capture longer temporal dynamics. We believe that both capturing longer temporal dynamics and the fact that our model is deeper have a high-impact on our model's performance.

The results depicted in Table 2 clearly show that our model outperforms the work by Trigeoris et al. [9] in both the arousal and valence dimensions for both the validation and test sets. More particularly, for the test set and for the arousal dimension with almost 3% absolute value and for the valence with a higher magnitude, i.e., 6%, absolute value.

	Arousal	Valence
Han et al. [16]	.729 (.785)	.309 (.364)
Han et al. [22]	.744 (.774)	.377 (.412)
Trigeorgis et al. [9]	.686 (.741)	.261 (.325)
Schmitt et al. [8]	.753 (.793)	.430 (.550)
Proposed	**.787** (.815)	**.440** (.502)

Table 2. Performance comparison (w.r.t ρ_c) between the proposed method and other state-of-the-art methods. In parenthesis are the performance obtained on the validation set.

These results indicate the effectiveness of our model, and as a consequence the effectiveness of our methodology in constructing DNNs.

We also compare our model with studies that utilise hand-crafted features for the emotion recognition task and produce state-of-the-art results on the RECOLA database. Table 2 depicts the best results of the studies with respect to the arousal dimension, as it can be more easily predicted from speech, and hence provide better results than the prediction of the valence dimension. In particular, Schmitt et al. [8] method produces the best results among the others, in both the arousal and valence dimensions. However, our model outperforms their method for both of the continuous affect dimensions. More specifically, for the prediction of the valence and arousal dimensions our model outperforms Schmitt et al. method with 1% and 3% absolute value, respectively. We should also mention that Schmitt et al. method uses hand-crafted features such as MFCCs, while our model was trained end-to-end with the architecture inspired by the creation of hand-crafted features.

A very recent study that utilised the full RECOLA dataset, as in our study but for multimodal input (audiovisual) was proposed by Tzirakis et al. [12]. The results obtained on the test set by the authors when utilising both the training and validation set for their multimodal model is .789, only 0.2% absolute value difference with our model that uses only the audio input and was trained only on the training set. For completeness, we mention that in the valence their method performance was 0.691 which is much higher than ours. This was expected as their model utilises also the video input which has been shown to more easily predict the valence dimension.

5.3. Depicting Arousal and Valence Predictions

To illustrate the effectiveness of our model, we choose to show the results obtained in a test video along with the ground truth. Figure 2 depicts the results. As can be observed our model can fit quite accurately the ground truth on the arousal dimension. However, this is not the case for the valence dimension.

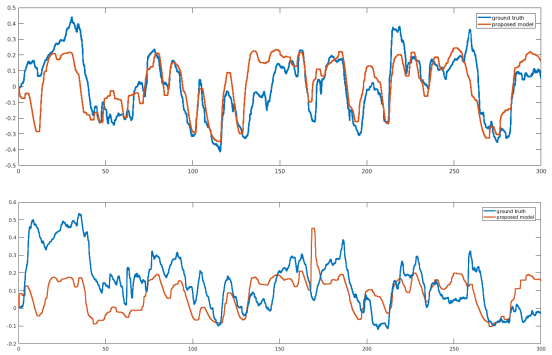

Fig. 2. Results obtained for a test subject for the arousal (Top) and valence (Bottom) dimensions.

6. CONCLUSION

In this paper, we propose a new convolution recurrent neural network structure for end-to-end speech emotion recognition. The proposed model achieve state-of-the-art results, with respect to the concordance correlation coefficient for both arousal and valence in comparison to previous studies which utilised the RECOLA database. Furthermore, we show the relationship between kernel and pooling size of the 1-d layers of our model, and window and step size for conventional audio features like MFCCs.

In future work, we will try deeper CNN models for audio analysis, utilising larger databases. We believe that we can acquire better performance for different audio analysis task using the raw signal. However, we still need to follow the basic idea for kernel size and pooling size when designing our model.

7. REFERENCES

[1] J. Russell, "A circular model of affect," *of Personality and Social Psychology*, pp. 1161–1178, 1980.

[2] 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 *ISCA*, 2008, pp. 597–600.

[3] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *ICASSP*, 2011, pp. 5688–5691.

[4] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *ICASSP*. IEEE, 2014, pp. 6964–6968.

[5] T. N Sainath, R. J Weiss, A. Senior, K. W Wilson, and O. Vinyals, "Learning the speech front-end with raw

waveform cldnns," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] T. N Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *ICASSP*. IEEE, 2015, pp. 4580–4584.

[7] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *ICASSP*. IEEE, 2017, pp. 421–425.

[8] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," *INTERSPEECH*, pp. 495–499, 2016.

[9] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*. IEEE, 2016, pp. 5200–5204.

[10] Michael Neumann and Ngoc Thang Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[11] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *Transactions on Affective Computing*, pp. 190–202, 2016.

[12] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[13] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *FG*. IEEE, 2013, pp. 1–8.

[14] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "Avec 2015: The 5th international audio/visual emotion challenge and workshop," in *MM*. ACM, 2015, pp. 1335–1336.

[15] M. Valstar, J. Gratch, B. 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 *AVEC Workshop*. ACM, 2016, pp. 3–10.

[16] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *ICASSP*. IEEE, 2017, pp. 2367–2371.

[17] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

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

[19] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels, "Ensemble methods for continuous affect recognition: Multimodality, temporality, and challenges," in *AVEC*, 2015, pp. 9–16.

[20] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels, "Ensemble methods for continuous affect recognition: Multi-modality, temporality, and challenges," in *AVEC Workshop*. ACM, 2015, pp. 9–16.

[21] 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, 2015.

[22] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in *ICASSP*. IEEE, 2017, pp. 5005–5009.