24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

End-To-End Speech Emotion Recognition Based on Time and Frequency Information Using Deep Neural Networks

Ali Bakhshi[1] and **Aaron S.W. Wong**[2] and **Stephan Chalup**[3]

Abstract. We propose a speech emotion recognition system based on deep neural networks, operating on raw speech data in an end-to-end manner to predict continuous emotions in arousal-valence space. The model is trained using time and frequency information of speech recordings of the publicly available part of the multi-modal RECOLA database. We use the Concordance Correlation Coefficient (CCC) as it was proposed by the Audio-Visual Emotion Challenges to measure the similarity between the network prediction and gold-standard. The CCC prediction results of our model outperform the results achieved by other state-of-the-art end-to-end models. The innovative aspect of our study is an end-to-end approach to using data that previously was mostly used by approaches involving combinations of pre-processing or post-processing. Our study used only a small subset of the RECOLA dataset and obtained better results than previous studies that used the full dataset.

1 INTRODUCTION

Emotions play an important role in human communication and can be observed in different channels such as speech and facial expressions. In fact, affective information is a fundamental component of human and machine communication [30]. Combination of different modalities such as physiological signals [20], audio [48, 26], video [12], hand gestures [24], and body movement [6] leads to many advancements in the field of affect recognition such as safe driving, security, and mobile health [22]. Automatic affect recognition systems can help to provide natural interfaces between humans and machines, whereby affect processing systems are designed to recognize and interpret human emotions and help the machine respond to perceived emotions. Understanding emotional states through speech signals as the fastest communication tool between humans has been investigated by many researchers in the past [8, 37, 17, 25, 38, 27, 16, 10, 39].

Applications of emotion recognition can be found in a range of areas. For instance, emotional states have been used for improving the learning processes and increasing the performance of e-learning systems [47]. Web designers and service providers try to modify and display the layout, content, and ads according to users' profiles that can be affected by their emotional states [15]. Speech emotion recognition systems can be used in call centers, where the aim is to improve the quality of call answering systems by giving appropriate feedback taking the emotional state of callers into account [2]. In the video games industry, that belongs to the wide range of entertainment applications, designers work on incorporating affective states and dreams of players into the gameplay [15]. Speech emotion recognition also finds application for violence detection systems in public places [1].

With the success of Deep Neural Networks (DNN) in solving complex machine learning problems in recent years, research in medical science, psychology, big data analysis, speech and speaker recognition, and paralinguistic problems tried to find improved solutions based on DNNs. Numerous researchers in academia, but also in industry are interested in this active area of research, and it has resulted in several groundbreaking improvements for many practical problems such as computer vision, human-machine interaction, speech, and natural language processing [29].

To date, most of studies on speech emotion recognition have utilized common hand-crafted features such as Mel-frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction (PLP) coefficients, and supra-segmental [32] for emotion classification. To this end, following extraction of these features referred to as low-level descriptors (LLDs), by applying some statistical functions such as mean, variance, and skewness to each of these LLDs, the high-level statistical functions (HSFs) are calculated. One functionality of the HSFs is describing the temporal variations of different LLDs that contain emotional information [23]. However, recent research methodologies have been directed towards end-to-end optimization, which only requires a minimum of prior human knowledge about the task and the data. The networks will eventually learn an intermediate representation of unprocessed input or raw data, that is necessary for real-world applications [9]. In the speech emotion recognition domain, limited research has been completed on end-to-end emotion recognition tasks [43, 44, 45, 40].

Inspired by the advancements in DNNs, we present an automatic affect recognition system that uses speech signals in an end-to-end manner. We introduce a deep Conv-RNN architecture based Gated Recurrent Units (GRU) [5] to extract audio features as well as to perform a regression task. Lin's Concordance Correlation Coefficient (ρ_c) [18] introduced by the Audio-Visual Emotion Challenges (AVECs) is used as similarity measurement tool between the network prediction and gold-standard, derived from the annotations of a dimensional model of emotion.

In our proposed model, the raw audio data is considered as the main input to the model. In addition, as a form of data augmentation, the FFT of the input signal is fed to the model to improve its performance.

The Remote Collaborative and Affective interactions (RECOLA) database [34] was used to show the performance of the proposed audio model. As the complete database is not publicly available, the model has been trained and validated on the part of the database that

[1] The University of Newcastle, Australia, email: ali.bakhshi@uon.edu.au
[2] 4Tel Pty. Ltd., Australia, email: awong@4tel.com.au
[3] The University of Newcastle, Callaghan NSW 2308, Australia, email: stephan.chalup@newcastle.edu.au

is publicly available. This database is recorded for the study of socio-affective behaviors from multimodal data [33]. This dataset has been used as the benchmark in AVEC challenges, and hence allows us to compare our model with other state-of-the-art models that use the same data.

The remainder of the paper organized as follows: Section 2 discusses related work on the speech emotion recognition systems. The contribution of this paper is presented in Section 3. Section 4 introduces the proposed audio model. Section 5 describes the dataset. Section 6 presents the experimental results, and Section 7 concludes the paper.

2 RELATED WORK

There exist sufficiently comprehensive surveys on speech emotion recognition (SER) efforts using traditional machine learning algorithms before 2011 [8, 28, 46]. Hence, in this section, we focus on discussing related work in speech emotion recognition with DNNs from 2011 onwards. The methodology of automatic emotion recognition was highly affected by DNNs. Different variants of DNN architectures have been used in emotion recognition tasks such as speech emotion recognition [19, 43, 13, 49, 9], facial emotion recognition [48], and multimodal emotion recognition [44]. Since this paper investigates speech signals, recent advancements on speech emotion recognition using DNNs are briefly reviewed in the remainder of this section.

With the introduction of DNNs a gradual integration of DNNs applied to the task of speech emotion recognition could be observed. Li et al. [19] proposed a hybrid DNN-Hidden Markov Model (HMM) using Restricted Boltzmann Machine (RBM) and trained it on MFCCs extracted from eNTRFACE'05 and Berlin databases for Speech Emotion Recognition (SER) in an unsupervised manner. The obtained results were superior when compared with traditional Gaussian Mixture Models, shallow-NN-HMMs with two layers, and Multi-Layer Perceptron HMMs (MLP-HMMs), testifying the suitability of DNNs for this task. Mao et al. [21] used a ConvNet architecture that learned affect-salient features using sparse autoencoders of speech spectrograms. Mao et al. employed a variant of the sparse autoencoder using unlabelled samples to learn local invariant features (LIF). These LIFs were then used as an input to the feature extractor, outperforming traditional speech emotion recognition features. Mao et al. showed the performance of the proposed model on four benchmarks involving the Surrey Audio-Visual Expressed Emotion (SAVEE), the Berlin Emotional Database (Emo-DB), the Danish Emotional Speech database (DES), and the Mandarin Emotional Speech database (MES). Jaebok et al. [13] used Multi-Task Learning (MTL) and gender and naturalness as auxiliary tasks in DNN to enhance emotion model generalization. In their DNN, they composed two hidden layers with 256 cells for LSTM-MTL for modeling temporal dynamics of emotion and three hidden layers of 256 nodes for DNN-MTL. In recent work, Zhao et al. [49] proposed deep 1-D and 2-D CNN-LSTM networks for speech emotion classification on the Berlin emotion database [3] and on the IEMOCAP database [4]. Zhao et al. used two different architectures, 1-D CNN-LSTM, and 2-D CNN-LSTM, that, both comprised four local feature learning blocks and one LSTM layer. Using the 2-D CNN-LSTM network, they achieved 95.33% and 95.89% classification accuracy on the Berlin database for speaker-dependent and speaker-independent tasks, respectively. Zhao et al. achieved recognition accuracies of 89.16% and 52.14% on the IEMOCAP database for speaker-dependent and speaker-independent tasks, respectively.

Tatinati et al. [40] proposed an end-to-end speech emotion recognition system using multi-scale convolution neural networks (MCNN) to detect features at various time scales and frequencies from raw speech signals. They showed that this tunable convolution network on the SAVEE emotion database improves the performance of the emotion recognition system compared to existing methods. In another study, Trigeorgis et al. [43] introduced an end-to-end speech emotion recognition system based on a deep convolutional recurrent network using the whole RECOLA dataset. They used the convolutional layers to extract features from raw audio data, and using Bidirectional Long Short-Term Memory (BiLSTM) layers they captured the temporal dependencies between the sequences of the audio frames. Also, they performed some pre-processing on the input and performed many post-processing steps such as median filtering, centering, time-shifting, and scaling to improve the prediction of the network, especially in the validation phase. More recent work [45] proposed a deeper network for continuous emotion recognition from the speech signals in the whole RECOLA dataset. They used three convolutional layers for feature extraction from the raw speech signal that each convolution layer followed by a max-pooling layer. The authors select the stride of each pooling layer such that the rate of overlap between the kernel size and stride of the pooling layer remains lower than 0.5. Also, they used two stacked LSTM layers on top of the convolutional and pooling layers, to consider the contextual information of the input data. Moreover, they used the same pre-processing and post-processing steps that were used in [43] to improve the prediction of the network. Most of the studies covered by literature for speech emotion recognition utilized hand-crafted audio features, and those that used raw signals, performed either pre-processing on the raw audio data or a chain of post-processing steps on the prediction of the network or both. In the present study, we propose a model trained in an end-to-end manner, with minimum pre-processing on speech data and without post-processing steps applied to the prediction of the network.

3 CONTRIBUTION OF THIS WORK

The main contribution of the present study is to use a relatively small set of data for training a deep neural network model on the time and frequency domain information of the raw audio data in an end-to-end manner. To our knowledge, this is the first work in literature that uses such a small set of raw audio signals as the input to the network, where the raw input is only complemented by the FFT of the signal. No post-processing on the raw prediction of the network is performed. All previous end-to-end learning models applied some pre-processing on the input data and a chain of post-processing steps on the prediction of the network [43, 44, 45]. The proposed model that uses the information of the time and the frequency domains shows very good performance for speech emotion recognition tasks, especially for the case of the arousal state. In fact, adding the information of the frequency domain acts as a data augmentation method to reduce the over-fitting problem caused by the low volume of the data during the training phase. The other point to mention about our neural network model is that it was trained using a similar number of CNN layers as the state-of-the-art end-to-end emotion recognition model of Tzirakis et al. [45] but used a smaller subset of the RECOLA dataset that was used by them for training. In all other end-to-end studies that use the RECOLA dataset, the authors have used the data of 16, 15, and 15 subjects for training, validation, and testing, respectively. In this paper, since the whole of the RECOLA dataset is not publicly available, we applied k-fold cross-validation

and utilized the data of 9 subjects for training and validating, and the data of separate 9 subjects for testing.

4 PROPOSED MODEL

In this study, we propose a Conv-BiGRU network as a feature extractor and predictor, to detect emotional states from speech signal data. Taking into account the characteristics of the time and frequency information in raw speech signals, we designed the deep speech emotion recognition model shown in figure 1. The design of our model and the associated experiments took into account experience from a series of pilot experiments where we tested different combinations of batch sizes, sequence lengths, and numbers of hidden layers. The final model comprises several stages where the different parts can be described as follows:

Stage 1: Two different inputs, the raw audio signal in the time-domain (upper stream in figure 1) and frequency-domain (lower stream in figure 1), with a specific time sequence length, are fed to the network. As illustrated in figure 1, there are three convolutional layers that extract features at the sampling rates of 16 kHz, 8 kHz, and 4 kHz. In the 16 kHz sampling rate of the input signal, each 1-sec sequence of the input signal corresponds to a 16000-dimensional vector.

Stage 2: Each input passes through a convolutional layer (Conv Layer 1) with a filter size of 40 and a kernel size of 16 to extract the finer-scale features of the speech signal.

Stage 3: In this stage, two max-pooling layers with different stride sizes are applied to the time and frequency domain features. A max-pooling layer with a stride of 2 (M-P1), decreases the frame rate of the data in the time domain. The max-pooling layer with a stride of 2 can be used as a down-sampling method to convert the input signal from 16 kHz to 8 kHz. In the frequency domain, the pooling layer with a stride of 8 (M-P) is applied across the channel to decrease the dimensionality of the data. In Stage 8, the output of the pooling layer in the frequency domain will be added to the final feature vector that is used as the input of BiGRU.

Stage 4: The second convolutional layer (Conv Layer 2) extracts the long term characteristic features of the speech signal. To this purpose, a convolution layer with a kernel size of 512 and a feature map size of 20 is applied to the input signal in the time domain.

Stage 5: In the time domain, the second pooling layer (M-P2) with a stride of 2 is applied to down-sample the signal to 4 kHz. As the telephony range of the human voice is below 4 kHz, considering the information extracted from this range can improve the performance of the network.

Stage 6: The third convolutional layer (Conv Layer 3) with a channel size of 10 and a kernel size of 1024 is applied across the time domain to extract more long term characteristics and higher-level abstractions of the input speech signal.

Stage 7: The final max-pooling layer (M-P3) with a stride of 8 is applied across the channel to reduce the dimensionality of the data.

Stage 8: The output features of the third max-pooling layer (M-P3) in the time domain and the output of the first pooling layer in the frequency domain (M-P) create the final feature vector that is used as the input of the BiGRU.

Stage 9: A two stacked-layer Bidirectional Gate Recurrent Unit (BiGRU) is used on top of the convolution and max-pooling layers to keep the temporal dependency between the samples in a sequence of audio frames and capturing the contextual information of the input data. When the training set is small, the GRU usually shows better results than LSTM.

Stage 10: The last BiGRU layer is followed by a linear fully connected layer with two outputs, corresponding to the arousal and valence dimensions.

Additional notes and details of the model:

- The first convolutional layer (Conv Layer 1) is followed by a ReLU function.
- All convolutional layers are followed by a batch normalization term where a momentum of 0.1 has been added.
- In several stages we employed max-pooling as a down-sampling method. Although resampling a signal from one frequency to a higher or lower frequency is a complicated process, in this case, the pooling layer operates like a down-sampling method. To understand the performance of the max-pooling layers for down-sampling we ran a number of tests: After the first convolutional layer, both, a down-sampling method and a max-pooling layer with a stride of 2 have been used in parallel to down-sample the data to 8 kHz. The CCC that was calculated between the outputs of the two down-sampling methods was above 0.95 indicating a high correlation between the outputs of the two methods. This means both methods perform similar in our situation but max-pooling is faster and simpler.
- Dropout regularization terms have been added in the time and frequency domain streams after each convolutional layer [41] to prevent over-fitting of the model. The dropout probability was set to 0.5 for the first convolutional layer (Conv Layer 1), and 0.75 for other convolutional layers (Conv Layers 2 and 3).
- The recurrent GRU layers address the sequence nature of the audio data and capture any long-term dependencies in the signal stream. The input feature size of the GRU layer is 320 in all experiments to match the structure of available data of the RECOLA database.

5 DATASET

Our study used the Remote Collaborative and Affective (RECOLA) database introduced by Ringeval et al. [34] as benchmark data for emotion recognition. This database consists of 9.5 hours of recordings of 46 French-speaking participants in four modalities; audio, visual, and two physiological (electrocardiogram (ECG) and electro-dermal activity (EDA)). The annotation was performed by six French-speaking assistants for each 40 msec interval within the first five minutes of interaction. This has resulted in 7500 values for arousal and valence states for each 5-minute video clip. The RECOLA provider team policies and the consent matter permitted only the data for half of the participants to be available in public. Consequently our study was restricted to use only that half of the data. In the RECOLA dataset, the time-continuous prediction of spontaneous emotions (arousal and valence) is considered. In the present study, the part of the RECOLA dataset that was used in the AVEC challenge has been utilized to make the performance of our network more comparable with a state-of-the-art model. This part of the RECOLA data includes the records of 27 subjects that are divided equally into 9 subjects for training, 9 subjects for validation, and 9 subjects for testing. Since the labels of the test data are only available for participants of the AVEC challenge, we used 10-fold cross-validation, where we used the data of 9 subjects for training and validating, and the data of separate 9 subjects for testing.

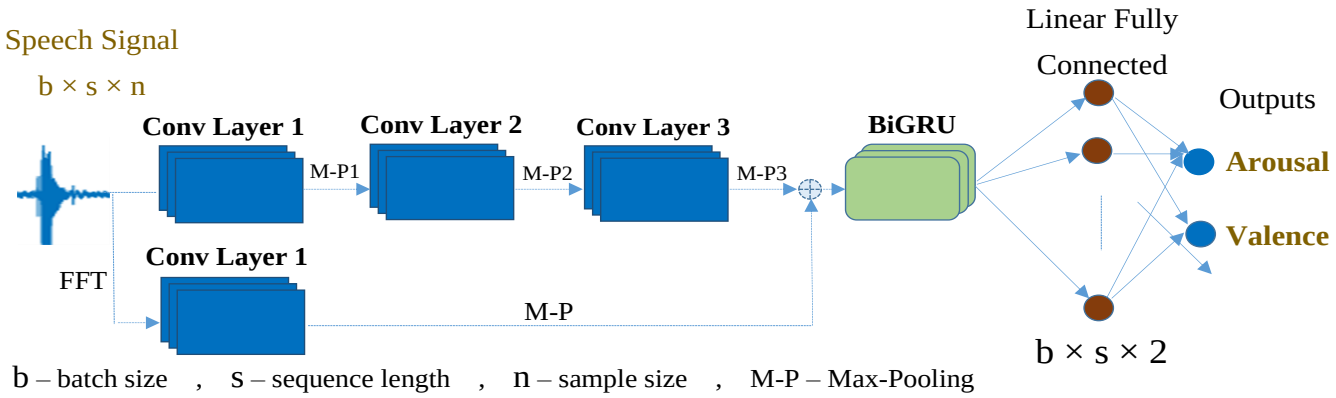

Figure 1. Proposed network architecture

6 EXPERIMENTS AND RESULTS

6.1 Objective function

In our experiments we use the Concordance Correlation Coefficient (ρ_c) [18] to evaluate the performance of our proposed emotion recognition model on speech signals extracted from video clips in the RECOLA database. Unlike most previous work that minimizes the MSE during training the network and then evaluates it using ρ_c [32, 31], we trained and evaluated our model using this metric (ρ_c) directly. That is, $L_c = 1 - \rho_c$ is used as an objective function where ρ_c is fully differentiable and can be integrated into gradient descent. Hence the cost function L_c that we used in the training and the evaluation phases can be described as follows:

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

where $\mu_x = E(x)$, $\mu_y = E(y)$, $\sigma_x^2 = var(x)$, $\sigma_y^2 = var(y)$, and $\sigma_{xy}^2 = cov(x, y)$. To minimize L_c (or maximize ρ_c), the gradients of L_c with respect to the last layer are backpropagated, where

$$\frac{\partial L_c}{\partial x} \propto 2\frac{\sigma_{xy}^2(x - \mu_y)}{\psi^2} + \frac{\mu_y - y}{\psi}, \tag{2}$$

where $\psi = \sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2$ and all vector operations are performed element-wise [44]. Since the mean and variance of x have to be estimated as part of this approach, the gradient described in equation (2) can only be used at batch mode [11].

6.2 Experimental setup

In this paper, various combinations of batch sizes and time-sequence lengths have been considered for evaluating the performance of the proposed model. We conducted several experiments where different segmentations of the raw speech signal were used as input: Each experiment used a fixed segment length. The segment length of the different experiments were 2, 4, 6, .. , 20 sec. At a 16 kHz sampling rate, each 1 sec corresponds to a 16000-dimensional vector. Hence, the 2, 4, 6, .. , 20 sec segments correspond to 32000 to 320000 dimensional input vectors.

In the training phase, the RMSProp optimizer [42] was used because it performed better in our pilot tests than other optimizers such as Adam [14], SGD [36], and Adagard [7]. A fixed learning rate of 10^{-4} with a learning rate decay of 10^{-5} was used throughout all our experiments.

6.3 Experimental results

The prediction performance of the model for different time sequence lengths and batch size 25 on the test set is shown in table 1. In this table, the prediction values for arousal and valence are the average prediction results of several runs, and the value in parenthesis is the standard deviation from the mean.

Table 1. The table shows the prediction performance of the proposed model for different input time sequence lengths and batch size 25. The best results are highlighted in bold font.

Sequence Length (sec)	Arousal	Valence
2	0.398 (0.0009)	0.142 (0.0016)
4	0.496 (0.0015)	0.174 (0.0034)
6	0.548 (0.0011)	0.237 (0.0019)
8	0.646 (0.0010)	**0.266** (0.0025)
10	0.622 (0.0038)	0.246 (0.0038)
12	0.635 (0.0047)	0.243 (0.0036)
14	0.649 (0.0006)	0.229 (0.0022)
16	**0.650** (0.0106)	0.209 (0.0052)
18	0.629 (0.0204)	0.181 (0.0477)
20	0.624 (0.0325)	0.176 (0.0150)

The values in parenthesis indicate the standard deviation.

As can be seen from table 1, for the Arousal dimension, the prediction performances of the network for time sequence lengths between 8 sec to 16 sec are very similar and the best performance was achieved for the sequence length of 16 sec. In case of the Valence dimension, the best prediction performance was achieved for 8 sec long input time sequences. Similar as in the case of the Arousal dimension, the prediction performances of the network for the Valence dimension for input time sequence lengths between 8 sec to 16 sec are very similar. If input time sequences become too short towards 2 sec or too long towards 20 sec the performance of the model drops for both input dimensions. The prediction results in table 1 indicate that input time sequence lengths above 8 sec and below 16 sec represent for this data the best time window size to recognize and capture

important emotional context and take the continuous nature of the emotional states during these time intervals into account.

To investigate the effect of selecting different batch sizes on the prediction performance of the network, training runs with batch sizes of 10, 25, and 50 were performed. The results are shown in table 2. As can be inferred from table 2, it is hard to specify a general best value for the batch size. But if we want to determine a suitable batch size value for the Arousal and the Valence dimension, the combination of a batch size of 50 with a time sequence length of 8 sec can be a good selection. Table 2 also shows that for each of the two input dimensions the prediction performances of the model for sequence lengths above 8 sec are very similar. The prediction results of table 2 are summarized and visualized in figure 2.

Table 2. The prediction of the proposed model for different batch sizes.

	Batch Size 10		Batch Size 25		Batch Size 50	
Seq_Length	Ar	Val	Ar	Val	Ar	Val
2	0.324	0.114	0.398	0.142	0.434	0.157
4	0.419	0.148	0.496	0.174	0.494	0.198
6	0.501	0.195	0.548	0.237	0.554	0.289
8	0.585	0.211	0.646	0.266	**0.660**	**0.314**
10	0.578	0.198	0.622	0.246	0.636	0.312
12	0.633	0.177	0.635	0.243	0.586	0.284
14	0.608	0.209	0.649	0.229	0.611	0.220
16	0.610	0.216	0.650	0.209	0.639	0.211
18	0.630	0.192	0.629	0.181	0.624	0.223
20	0.621	0.193	0.624	0.176	0.612	0.228

Seq_Length: Sequence Length Ar: Arousal Val: Valence

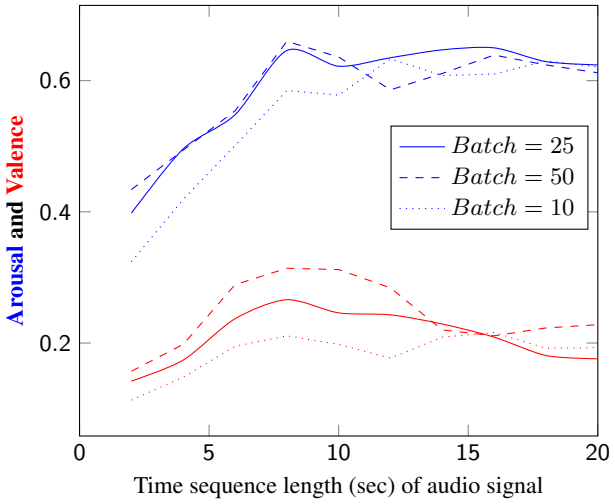

Figure 2. Comparison of the prediction performances of the proposed model when using different batch sizes (blue lines: Arousal, red lines: Valence)

Our results in table 2 confirm that the selection of a well-performing batch size highly depends on the type of input data and the nature of the problem. As can be seen in figure 2, the behavior of the model for prediction of the arousal state for batch sizes 25 and 50 is almost the same, but for prediction of the valence state batch size 50 performs better.

A direct comparison between the prediction of the proposed model and other state-of-the-art models on the RECOLA dataset was not possible, due to the different versions and combinations of training, validation, and test sets. Therefore, we restricted our comparative evaluation of the performance of other state-of-the-art end-to-end models to the same subset of the RECOLA dataset that we used in the present paper. Table 3 compares the prediction performances of these networks in terms of ρ_c for the Arousal and Valence dimensions. For all methods, a batch size of 50 and a time sequence length of 8 sec of the input signal has been used. The prediction results of the proposed model for the Arousal and Valence dimensions outperformed the results of the two other models that were used for comparison. Although our network does not predict the Valence dimension very well, the prediction results for the Arousal dimension are comparable with the models that used the whole RECOLA dataset for training, validating, and testing [43, 44].

Table 3. Prediction performance of the proposed model and related work on the RECOLA dataset

Method	Arousal	Valence
ConvBiLSTM [43]	0.559	0.128
ConvLSTM [45]	0.546	0.183
Proposed	0.660	0.314

6.4 Discussion

There are many pre-trained deep network models that are designed for two-dimensional data like images or video, but there are not so many trained models for one-dimensional data like speech. In one dimensional data and particularly in the speech signals, most of the useful information exists in a particular frequency range, usually below 8 kHz [35]. Therefore, selecting a suitable sampling rate for the input signal can strongly influence the operation of the deep networks. Besides, some hyper-parameters of the deep model, such as kernel size and the number of channels in the convolutional layers, as well as the stride of the pooling layers can be selected regarding the information available in various sampling rates of the speech signal. For example, selecting a suitable kernel size for the first convolutional layer is very important because using an inappropriate kernel size can cause undesired changes on the input features and structure, and thus it can considerably change the prediction of the network. We found in our experiments that selecting large values for the kernel size of the first convolution layer can corrupt the prediction ability of the network. Although the kernel sizes of other convolution layers are important, they have not the same significant impact on the prediction ability of the network as the kernel size of the first layer. We also found that a suitable batch size and time sequence length of the input data are among the effective parameters that most impact on the prediction of the network. According to our experiments for spoken data in the RECOLA dataset, the best batch value is 50, although the prediction results of the network for the batch size 25 are very close. Given the continuous nature of emotions over a period of time, considering a temporal sequence of speech signals as network inputs can improve model prediction because recurrent layers can capture the contextual information of the data sequences. In this study, different sequence lengths of input data were considered, and the best predictions for the time sequence are greater than 8 seconds. It is very difficult to select a particular time sequence length as it depends on many parameters, such as the structure of the data and batch size.

7 CONCLUSION

In this paper we proposed a speech emotion recognition system based on Conv-BiGRU layers that uses the raw audio data of the RECOLA dataset in an end-to-end manner. Our experiments addressed various combinations of time sequence lengths and batch sizes to understand the impact of these parameters on the prediction performance of the network better. By using the information of the time and the frequency domain, we designed a deep network that could be trained on a relatively small set of training data. In the experiments, our model outperformed state-of-the-art end-to-end models in the prediction of arousal and valence states on the available part of the RECOLA dataset.

ACKNOWLEDGEMENTS

Ali Bakhshi was supported by a UNIPRS PhD scholarship at The University of Newcastle, Australia. The authors are grateful to the UON HPC IT-team for their support and to the SEEC Experimental Deep Learning Supercomputing Lab for providing access to their DGX station. Some of the computational (and/or storage) resources used in this work were enabled by Intersect Australia Limited and partially subsidised by funding from the Australian Research Council through ARC LIEF Grants LE160100002 and LE170100032.

REFERENCES

[1] Marta Bautista-Duran, Joaquin Garcia-Gomez, Roberto Gil-Pita, Inma Mohino-Herranz, and Manuel Rosa-Zurera, 'Energy-efficient acoustic violence detector for smart cities', *International Journal of Computational Intelligence Systems*, **10**, 1298–1305, (2017).

[2] Felix Burkhardt, Jitendra Ajmera, Roman Englert, Joachim Stegmann, and Winslow Burleson, 'Detecting anger in automated voice portal dialogs', in *Ninth International Conference on Spoken Language Processing*, (2006).

[3] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, 'A database of german emotional speech', in *Ninth European Conference on Speech Communication and Technology*, (2005).

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, 'Iemocap: Interactive emotional dyadic motion capture database', *Language resources and evaluation*, **42**(4), 335, (2008).

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, 'Learning phrase representations using rnn encoder-decoder for statistical machine translation', *arXiv preprint arXiv:1406.1078*, (2014).

[6] Mohamed Daoudi, Stefano Berretti, Pietro Pala, Yvonne Delevoye, and Alberto Del Bimbo, 'Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices', in *International Conference on Image Analysis and Processing*, pp. 550–560. Springer, (2017).

[7] John Duchi, Elad Hazan, and Yoram Singer, 'Adaptive subgradient methods for online learning and stochastic optimization', *Journal of Machine Learning Research*, **12**(Jul), 2121–2159, (2011).

[8] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, 'Survey on speech emotion recognition: Features, classification schemes, and databases', *Pattern Recognition*, **44**(3), 572–587, (2011).

[9] Alex Graves and Navdeep Jaitly, 'Towards end-to-end speech recognition with recurrent neural networks', in *International Conference on Machine Learning*, pp. 1764–1772, (2014).

[10] Ali Harimi, Hasan Shaygan Fakhr, and Ali Bakhshi, 'Recognition of emotion using reconstructed phase space of speech', *Malaysian Journal of Computer Science*, **29**(4), 262–271, (2016).

[11] Markus Kächele, Patrick Thiam, Günther Palm, Friedhelm Schwenker, and Martin Schels, 'Ensemble methods for continuous affect recognition: Multi-modality, temporality, and challenges', in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pp. 9–16. ACM, (2015).

[12] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah, 'Video-based emotion recognition in the wild using deep transfer learning and score fusion', *Image and Vision Computing*, **65**, 66–75, (2017).

[13] Jaebok Kim, Gwenn Englebienne, Khiet P Truong, and Vanessa Evers, 'Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning', *arXiv preprint arXiv:1708.03920*, (2017).

[14] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*, (2014).

[15] Agata Kołakowska, Agnieszka Landowska, Mariusz Szwoch, Wioleta Szwoch, and Michal R Wrobel, 'Emotion recognition and its applications', in *Human-Computer Systems Interaction: Backgrounds and Applications 3*, 51–62, Springer, (2014).

[16] Shashidhar G Koolagudi and K Sreenivasa Rao, 'Emotion recognition from speech: a review', *International Journal of Speech Technology*, **15**(2), 99–117, (2012).

[17] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee, 'Emotion recognition by speech signals', in *Eighth European Conference on Speech Communication and Technology*, (2003).

[18] I Lawrence and Kuei Lin, 'A concordance correlation coefficient to evaluate reproducibility', *Biometrics*, 255–268, (1989).

[19] Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin, and Hichem Sahli, 'Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition', in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pp. 312–317. IEEE, (2013).

[20] Choubeila Maaoui and Alain Pruski, 'Emotion recognition through physiological signals for human-machine communication', in *Cutting Edge Robotics 2010*, InTech, (2010).

[21] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan, 'Learning salient features for speech emotion recognition using convolutional neural networks', *IEEE Transactions on Multimedia*, **16**(8), 2203–2213, (2014).

[22] Catherine Marechal, Dariusz Mikołajewski, Krzysztof Tyburek, Piotr Prokopowicz, Lamine Bougueroua, Corinne Ancourt, and Katarzyna Węgrzyn-Wolska, 'Survey on ai-based multimodal methods for emotion detection', in *High-Performance Modelling and Simulation for Big Data Applications*, 307–324, Springer, (2019).

[23] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, 'Automatic speech emotion recognition using recurrent neural networks with local attention', in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 2227–2231. IEEE, (2017).

[24] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari, 'Survey on emotional body gesture recognition', *arXiv preprint arXiv:1801.07481*, (2018).

[25] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva, 'Speech emotion recognition using hidden markov models', *Speech communication*, **41**(4), 603–623, (2003).

[26] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic, 'End-to-end audiovisual speech recognition', *CoRR*, **abs/1802.06424**, (2018).

[27] Valery A Petrushin, 'Emotion recognition in speech signal: experimental study, development, and application', in *Sixth International Conference on Spoken Language Processing*, (2000).

[28] Paolo Petta, Catherine Pelachaud, and Roddy Cowie, *Emotion-oriented systems: the HUMAINE handbook*, Springer, 2011.

[29] Dhanesh Ramachandram and Graham W Taylor, 'Deep multimodal learning: A survey on recent advances and trends', *IEEE Signal Processing Magazine*, **34**(6), 96–108, (2017).

[30] Fuji Ren, 'Affective information processing and recognizing human emotion', *Electronic notes in theoretical computer science*, **225**, 39–50, (2009).

[31] 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, (2015).

[32] 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 *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–8. ACM, (2015).

[33] Fabien Ringeval, Andreas Sonderegger, Basilio Noris, Aude Billard, Juergen Sauer, and Denis Lalanne, 'On the influence of emotional feedback on emotion awareness and gaze behavior', in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pp. 448–453. IEEE, (2013).

[34] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne, 'Introducing the recola multimodal corpus of remote collaborative and affective interactions', in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pp. 1–8. IEEE, (2013).

[35] Stuart Rosen and Peter Howell, *Signals and systems for speech and hearing*, volume 29, Brill, 2011.

[36] Sebastian Ruder, 'An overview of gradient descent optimization algorithms', *arXiv preprint arXiv:1609.04747*, (2016).

[37] Björn Schuller, Gerhard Rigoll, and Manfred Lang, 'Hidden markov model-based speech emotion recognition', in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 2, pp. II–1. IEEE, (2003).

[38] Björn Schuller, Gerhard Rigoll, and Manfred Lang, 'Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture', in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04).*, volume 1, pp. I–577. IEEE, (2004).

[39] Björn W Schuller, 'Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends', *Communications of the ACM*, **61**(5), 90–99, (2018).

[40] Tatinati Sivanagaraja, Mun Kit Ho, Andy WH Khong, and Yubo Wang, 'End-to-end speech emotion recognition using multi-scale convolution networks', in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 189–192. IEEE, (2017).

[41] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, 'Dropout: a simple way to prevent neural networks from overfitting', *The Journal of Machine Learning Research*, **15**(1), 1929–1958, (2014).

[42] Tijmen Tieleman and Geoffrey Hinton, 'Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude', *COURSERA: Neural networks for machine learning*, **4**(2), 26–31, (2012).

[43] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, 'Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network', in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 5200–5204. IEEE, (2016).

[44] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou, 'End-to-end multimodal emotion recognition using deep neural networks', *IEEE Journal of Selected Topics in Signal Processing*, **11**(8), 1301–1309, (2017).

[45] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller, 'End-to-end speech emotion recognition using deep neural networks', in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5089–5093. IEEE, (2018).

[46] Dimitrios Ververidis and Constantine Kotropoulos, 'Emotional speech recognition: Resources, features, and methods', *Speech Communication*, **48**(9), 1162–1181, (2006).

[47] Jiangqin Xu, Zhongqiang Huang, Minghui Shi, and Min Jiang, 'Emotion detection in e-learning using expectation-maximization deep spatial-temporal inference network', in *UK Workshop on Computational Intelligence*, pp. 245–252. Springer, (2017).

[48] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, Jingwei Yan, and Keyu Yan, 'A deep neural network-driven feature learning method for multi-view facial expression recognition', *IEEE Transactions on Multimedia*, **18**(12), 2528–2536, (2016).

[49] Jianfeng Zhao, Xia Mao, and Lijiang Chen, 'Speech emotion recognition using deep 1d & 2d cnn lstm networks', *Biomedical Signal Processing and Control*, **47**, 312–323, (2019).