




Towards Robust Deep Neural Networks for Affect and Depression Recognition from Speech

Alice Othmani¹ , Daoud Kadoch², Kamil Bentounes², Emna Rejaibi³, Romain Alfred⁴, and Abdenour Hadid⁵

¹ University of Paris-Est Créteil, Vitry sur Seine, France

alice.othmani@u-pec.fr

² Sorbonne University, Paris, France

³ INSAT, Tunis, Tunisie

⁴ ENSIIE, Évry, France

⁵ Polytechnic University of Hauts-de-France, Valenciennes, France

Abstract. Intelligent monitoring systems and affective computing applications have emerged in recent years to enhance healthcare. Examples of these applications include assessment of affective states such as Major Depressive Disorder (MDD). MDD describes the constant expression of certain emotions: negative emotions (low Valence) and lack of interest (low Arousal). High-performing intelligent systems would enhance MDD diagnosis in its early stages. In this paper, we present a new deep neural network architecture, called EmoAudioNet, for emotion and depression recognition from speech. Deep EmoAudioNet learns from the time-frequency representation of the audio signal and the visual representation of its spectrum of frequencies. Our model shows very promising results in predicting affect and depression. It works similarly or outperforms the state-of-the-art methods according to several evaluation metrics on RECOLA and on DAIC-WOZ datasets in predicting arousal, valence, and depression. Code of EmoAudioNet is publicly available on GitHub: <https://github.com/AliceOTHMANI/EmoAudioNet>.

Keywords: Emotional Intelligence · Socio-affective computing · Depression recognition · Speech emotion recognition · Healthcare application · Deep learning.

1 Introduction

Artificial Emotional Intelligence (EI) or affective computing has attracted increasing attention from the scientific community. Affective computing consists of endowing machines with the ability to recognize, interpret, process and simulate human affects. Giving machines skills of emotional intelligence is an important key to enhance healthcare and further boost the medical assessment of several mental disorders.

© Springer Nature Switzerland AG 2021

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12662, pp. 5–19, 2021.

https://doi.org/10.1007/978-3-030-68790-8_1

Affect describes the experience of a human’s emotion resulting from an interaction with stimuli. Humans express an affect through facial, vocal, or gestural behaviors. A happy or angry person will typically speak louder and faster, with strong frequencies, while a sad or bored person will speak slower with low frequencies. Emotional arousal and valence are the two main dimensional affects used to describe emotions. Valence describes the level of pleasantness, while arousal describes the intensity of excitement. A final method for measuring a user’s affective state is to ask questions and to identify emotions during an interaction. Several post-interaction questionnaires exist for measuring affective states like the Patient Health Questionnaire 9 (PHQ-9) for depression recognition and assessment. The PHQ is a self report questionnaire of nine clinical questions where a score ranging from 0 to 23 is assigned to describe Major Depressive Disorder (MDD) severity level. MDD is a mental disease which affects more than 300 million people in the world [1], *i.e.*, 3% of the worldwide population. The psychiatric taxonomy classifies MDD among the low moods [2], *i.e.*, a condition characterised by a tiredness and a global physical, intellectual, social and emotional slow-down. In this way, the speech of depressive subjects is slowed, the pauses between two speakings are lengthened and the tone of the voice (prosody) is more monotonous.

In this paper, a new deep neural networks architecture, called EmoAudioNet, is proposed and evaluated for real-life affect and depression recognition from speech. The remainder of this article is organised as follows. Section 2 introduces related works with affect and depression recognition from speech. Section 3 introduces the motivations behind this work. Section 4 describes the details of the overall proposed method. Section 5 describes the entire experiments and the extensive experimental results. Finally, the conclusion and future work are presented in Sect. 6.

2 Related Work

Several approaches are reported in the literature for affect and depression recognition from speech. These methods can be generally categorized into two groups: hand-crafted features-based approaches and deep learning-based approaches.

2.1 Handcrafted Features-Based Approaches

In this family of approaches, there are two main steps: feature extraction and classification. An overview of handcrafted features-based approaches for affect and depression assessment from speech is presented in Table 1.

Handcrafted Features. Acoustic Low-Level Descriptors (LLD) are extracted from the audio signal. These LLD are grouped into four main categories: the **spectral LLD** (Harmonic Model and Phase Distortion Mean (HMPDM0-24), etc.), the **cepstral LLD** (Mel-Frequency Cepstral Coefficients (MFCC) [3, 13], etc.), the **prosodic LLD** (Formants [21], etc.), and the **voice quality LLD**

(Jitter, and Shimmer [14], etc.). A set of statistical features are also calculated (max, min, variance and standard deviation of LLD [4, 12]). Low *et al.* [16] propose the experimentation of the Teager Energy Operator (TEO) based features.

A comparison of the performances of the prosodic, spectral, glottal (voice quality), and TEO features for depression recognition is realized in [16] and it demonstrates that the different features have similar accuracies. The fusion of the prosodic LLD and the glottal LLD based models seems to not significantly improve the results, or decreased them. However, the addition of the TEO features improves the performances up to +31,35% for depressive male.

Classification of Handcrafted Features. Comparative analysis of the performances of several classifiers in depression assessment and prediction indicate that the use of an hybrid classifier using Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) model gave the best overall classification results [6, 16]. Different fusion methods, namely feature, score and decision fusion have been also investigated in [6] and it has been demonstrated that: first, amongst the fusion methods, score fusion performed better when combined with GMM, HFS and MLP classifiers. Second, decision fusion worked best for SVM (both for raw data and GMM models) and finally, feature fusion exhibited weak performance compared to other fusion methods.

2.2 Deep Learning-Based Approaches

Recently, approaches based on deep learning have been proposed [8, 23–30]. Several handcrafted features are extracted from the audio signals and fed to the deep neural networks, except in Jain [27] where only the MFCC are considered. In other approaches, raw audio signals are fed to deep neural networks [19]. An overview of deep learning-based methods for affect and depression assessment from speech is presented in Table 2.

Several deep neural networks have been proposed. Some deep architectures are based on feed-forward neural networks [11, 20, 24], some others are based on convolutional neural networks such as [27] and [8] whereas some others are based on recurrent neural networks such as [13] and [23]. A comparative study [25] of some neural networks, BLSTM-MIL, BLSTM-RNN, BLSTM-CNN, CNN, DNN-MIL and DNN, demonstrates that the BLSTM-MIL outperforms the other studied architectures. Whereas, in Jain [27], the Capsule Network is demonstrated as the most efficient architecture, compared to the BLSTM with Attention mechanism, CNN and LSTM-RNN. For the assessment of the level of depression using the Patient Health Questionnaire 8 (PHQ-8), Yang *et al.* [8] exerts a DCNN. To the best of our knowledge, their approach outperforms all the existing approaches on DAIC-WOZ dataset.

3 Motivations and Contributions

Short-time spectral analysis is the most common way to characterize the speech signal using MFCCs. However, audio signals in their time-frequency

Table 1. Overview of shallow learning based methods for affect and depression assessment from speech. (*) Results obtained over a group of females.

Ref	Features	Classification	Dataset	Metrics	Value
Valstar <i>et al.</i> [3]	prosodic + voice quality + spectral	SVM + grid search + random forest	DAIC-WOZ	F1-score Precision Recall RMSE (MAE)	0.410 (0.582) 0.267 (0.941) 0.889 (0.421) 7.78 (5.72)
Dhall <i>et al.</i> [14]	energy + spectral + voicing quality + duration features	non-linear chi-square kernel	AFEW 5.0	unavailable	unavailable
Ringeval <i>et al.</i> [4]	prosodic LLD + voice quality + spectral	random forest	SEWA	RMSE MAE	7.78 5.72
Haq <i>et al.</i> [15]	energy + prosodic + spectral + duration features	Sequential Forward Selection + Sequential Backward Selection + linear discriminant analysis + Gaussian classifier uses Bayes decision theory	Natural speech databases	Accuracy	66.5%
Jiang <i>et al.</i> [5]	MFCC + prosodic + spectral LLD + glottal features	ensemble logistic regression model for detecting depression E algorithm	hand-crafted dataset	Males accuracy Males sensitivity Males specificity	81.82%(70.19%*) 78.13%(79.25%*) 85.29%(70.59%*)
Low <i>et al.</i> [16]	teager energy operator based features	Gaussian mixture model + SVM	hand-crafted dataset	Males accuracy Males sensitivity Males specificity	86.64%(78.87%*) 80.83%(80.64%*) 92.45%(77.27%*)
Alghowinem <i>et al.</i> [6]	energy + formants + glottal features + intensity + MFCC + prosodic + spectral + voice quality	Gaussian mixture model + SVM + decision fusion	hand-crafted dataset	Accuracy	91.67%
Valstar <i>et al.</i> [7]	duration features+energy local min/max related functionals+spectral+voicing quality	correlation based feature selection + SVR + 5-fold cross-validation loop	AViD-Corpus	RMSE MAE	14.12 10.35
Valstar <i>et al.</i> [17]	duration features+energy local min/max related functionals+spectral+voicing quality	SVR	AVEC2014	RMSE MAE	11.521 8.934
Cummins <i>et al.</i> [9]	MFCC + prosodic + spectral centroid	SVM	AVEC2013	Accuracy	82%
Lopez Otero <i>et al.</i> [10]	energy + MFCC + prosodic + spectral	SVR	AVDLC	RMSE (MAE)	8.88 (7.02)
Meng <i>et al.</i> [18]	spectral + energy + MFCC + functionals features + duration features	PLS regression	AVEC2013	RMSE MAE CORR	11.54 9.78 0.42

Table 2. Overview of deep learning based methods for affect and depression assessment from speech.

Ref	Features	Classification	Dataset	Metrics	Value
Yang <i>et al.</i> [8]	spectral LLD + cepstral LLD + prosodic LLD + voice quality LLD + statistical functionals + regression functionals	DCNN	DAIC-WOZ	Depressed female RMSE	4.590
				Depressed female MAE	3.589
				Not depressed female RMSE	2.864
				Not depressed female MAE	2.393
				Depressed male RMSE	1.802
				Depressed male MAE	1.690
				Not depressed male RMSE	2.827
				Not depressed male MAE	2.575
Al Hanai <i>et al.</i> [23]	spectral LLD + cepstral LLD + prosodic LLD + voice quality LLD + functionals	LSTM-RNN	DAIC	F1-score	0.67
				Precision	1.00
				Recall	0.50
				RMSE	10.03
Dham <i>et al.</i> [24]	prosodic LLD + voice quality LLD + functionals + BoTW	FF-NN	AVEC2016	RMSE	7.631
				MAE	6.2766
Salekin <i>et al.</i> [25]	spectral LLD + MFCC + functionals	NN2Vec + BLSTM-MIL	DAIC-WOZ	F1-score	0.8544
				Accuracy	96.7%
Yang <i>et al.</i> [26]	spectral LLD + cepstral LLD + prosodic LLD + voice quality LLD + functionals	DCNN-DNN	DAIC-WOZ	Female RMSE	5.669
				Female MAE	4.597
				Male RMSE	5.590
				Male MAE	5.107
Jain [27]	MFCC	Capsule Network	VCTK corpus	Accuracy	0.925
Chao <i>et al.</i> [28]	spectral LLD + cepstral LLD + prosodic LLD	LSTM-RNN	AVEC2014	<i>unavailable</i>	<i>unavailable</i>
Gupta <i>et al.</i> [29]	spectral LLD + cepstral LLD + prosodic LLD + voice quality LLD + functionals	DNN	AVID-Corpus	<i>unavailable</i>	<i>unavailable</i>
Kang <i>et al.</i> [30]	spectral LLD + prosodic LLD + articulatory features	DNN	AVEC2014	RMSE	7.37
				MAE	5.87
				Pearson's Product Moment Correlation coefficient	0.800
Tzirakis <i>et al.</i> [36]	raw signal	CNN and 2-layers LSTM	RECOLA	loss function based on CCC	.440(arousal) .787(valence)
Tzirakis <i>et al.</i> [19]	raw signal	CNN and LSTM	RECOLA	CCC	.686(arousal) .261(valence)
					Tzirakis <i>et al.</i> [22]

representations, often present interesting patterns in the visual domain [31]. The visual representation of the spectrum of frequencies of a signal using its spectrogram shows a set of specific repetitive patterns. Surprisingly and to the best of our knowledge, it has not been reported in the literature a deep neural network architecture that combines information from time, frequency and visual domains for emotion recognition.

The first contribution of this work is a new deep neural network architecture, called EmoAudioNet, that aggregate responses from a short-time spectral analysis and from time-frequency audio texture classification and that extract deep features representations in a learned embedding space. In a second contribution, we propose EmoAudioNet-based approach for instantaneous prediction of spontaneous and continuous emotions from speech. In particular, our specific contributions are as follows: (i) an automatic clinical depression recognition and assessment embedding network (ii) a small size two-stream CNNs to map audio data into two types of continuous emotional dimensions namely, arousal and valence and (iii) through experiments, it is shown that EmoAudioNet-based features outperforms the state-of-the art methods for predicting depression on DAIC-WOZ dataset and for predicting valence and arousal dimensions in terms of Pearson’s Coefficient Correlation (PCC).

Algorithm 1 EmoAudioNet embedding network.

Given two feature extractors f_θ and f_ϕ , number of training steps N .

for *iteration in range*(N) **do**

$(\mathbf{X}_{\text{wav}}, \mathbf{y}_{\text{wav}}) \leftarrow$ batch of input wav files and labels

$\mathbf{e}_{\text{Spec}} \leftarrow f_\theta(\mathbf{X}_{\text{wav}})$ Spectrogram features

$\mathbf{e}_{\text{MFCC}} \leftarrow f_\phi(\mathbf{X}_{\text{wav}})$ MFCC features

$\mathbf{f}_{\text{MFCCSpec}} \leftarrow [\mathbf{e}_{\text{MFCC}}, \mathbf{e}_{\text{Spec}}]$ Feature-level fusion

$\mathbf{p}_{\text{MFCCSpec}} \leftarrow f_\theta(\mathbf{f}_{\text{MFCCSpec}})$ Predict class probabilities

$L_{\text{MFCCSpec}} = \text{cross_entropy_loss}(\mathbf{p}_{\text{MFCCSpec}}, \mathbf{y}_{\text{wav}})$

 Obtain all gradients $\Delta_{\text{all}} = (\frac{\partial L}{\partial \theta}, \frac{\partial L}{\partial \phi})$

$(\theta, \phi, \theta) \leftarrow \text{ADAM}(\Delta_{\text{all}})$ Update feature extractor and output heads’ parameters simultaneously

end

4 Proposed Method

We seek to learn a deep audio representation that is trainable end-to-end for emotion recognition. To achieve that, we propose a novel deep neural network called EmoAudioNet, which performs low-level and high-level features extraction and aggregation function learning jointly (See Algorithm 1). Thus, the input audio signal is fed to a small size two-stream CNNs that outputs the final classification scores. A data augmentation step is considered to increase the amount of data by adding slightly modified copies of already existing data. The structure of EmoAudioNet presents three main parts as shown in Fig. 1: (i) An MFCC-based CNN, (ii) A spectrogram-based CNN and (iii) the aggregation of the responses of the MFCC-based and the spectrogram-based CNNs. In the following, more details about the three parts are given.

4.1 Data Augmentation

A data augmentation step is considered to overcome the problem of data scarcity by increasing the quantity of training data and also to improve the model’s robustness to noise. Two different types of audio augmentation techniques are performed: **(1) Adding noise:** mix the audio signal with random noise. Each mix z is generated using $z = x + \alpha \times rand(x)$ where x is the audio signal and α is the noise factor. In our experiments, $\alpha = 0.01, 0.02$ and 0.03 . **(2) Pitch Shifting:** lower the pitch of the audio sample by 3 values (in semitones): (0.5, 2 and 5).

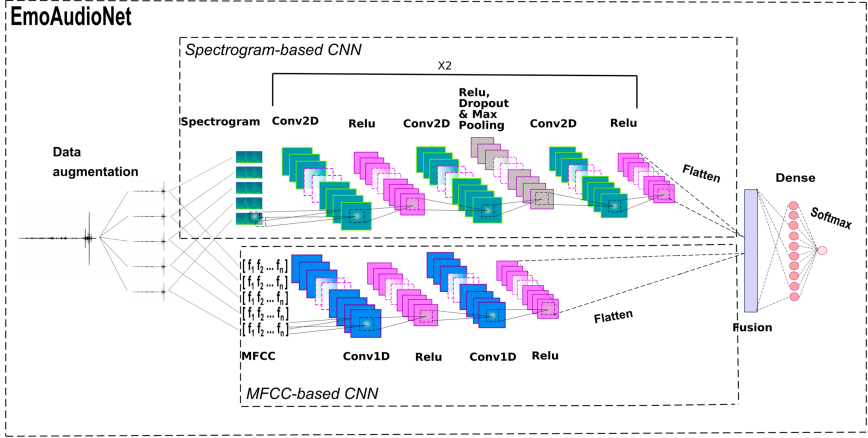


Fig. 1. The diagram of the proposed deep neural networks architecture called EmoAudioNet. The output layer is dense layer of size n neurones with a Softmax activation function. n is defined according to the task. When the task concerns binary depression classification, $n = 2$. When the task concerns depression severity level assessment, $n = 24$. While, $n = 10$ for arousal or valence prediction.

4.2 Spectrogram-Based CNN Stream

The spectrogram-based CNN presents low-level features descriptor followed by a high-level features descriptor. The Low-level features descriptor is the spectrogram of the input audio signal and it is computed as a sequence of Fast Fourier Transform (FFT) of windowed audio segments. The audio signal is split into 256 segments and the spectrum of each segment is computed. The Hamming window is applied to each segment. The spectrogram plot is a color image of $1900 \times 1200 \times 3$. The image is resized to $224 \times 224 \times 3$ before being fed to the High-level features descriptor. The high-Level features descriptor is a deep CNN, it takes as input the spectrogram of the audio signal. Its architecture, as shown in Fig. 1, is composed by two same blocks of layers. Each block is composed of a two-dimensional (2D) convolutional layer followed by a ReLU activation function, a second convolutional layer, a ReLU, a dropout and max pooling layer, a third convolutional layer and last ReLU.

4.3 MFCC-Based CNN Stream

The MFCC-based CNN presents also a low-level followed by high-level features descriptors (see Fig. 1). The low-level features descriptor is the MFCC features of the input audio. To extract them, the speech signal is first divided into frames by applying a Hamming windowing function of 2.5 s at fixed intervals of 500 ms. A cepstral feature vector is then generated and the Discrete Fourier Transform (DFT) is computed for each frame. Only the logarithm of the amplitude spectrum is retained. The spectrum is after smoothed and 24 spectral components into 44100 frequency bins are collected in the Mel frequency scale. The components of the Mel-spectral vectors calculated for each frame are highly correlated. Therefore, the Karhunen-Loeve (KL) transform is applied and is approximated by the Discrete Cosine Transform (DCT). Finally, 177 cepstral features are obtained for each frame. After the extraction of the MFCC features, they are fed to the high-Level features descriptor which is a small size CNN. To avoid overfitting problem, only two one-dimensional (1D) convolutional layers followed by a ReLU activation function each are performed.

4.4 Aggregation of the Spectrogram-Based and MFCC-Based Responses

Combining the responses of the two deep streams CNNs allows to study simultaneously the time-frequency representation and the texture-like time frequency representation of the audio signal. The output of the spectrogram-based CNN is a feature vector of size 1152, while the output of the MFCC-based CNN is a feature vector of size 2816. The responses of the two networks are concatenated and then fed to a fully connected layer in order to generate the label prediction of the emotion levels.

5 Experiments and Results

5.1 Datasets

Two publicly available datasets are used to evaluate the performances of EmoAudioNet:

Dataset for Affect Recognition Experiments: RECOLA dataset [32] is a multimodal corpus of affective interactions in French. 46 subjects participated to data recordings. Only 23 audio recordings of 5 min of interaction are made publicly available and used in our experiments. Participants engaged in a remote discussion according to a survival task and six annotators measured emotion continuously on two dimensions: valence and arousal.

Dataset for Depression Recognition and Assessment Experiments: DAIC-WOZ depression dataset [33] is introduced in the AVEC2017 challenge [4] and it provides audio recordings of clinical interviews of 189 participants.

Each recording is labeled by the PHQ-8 score and the PHQ-8 binary. The PHQ-8 score defines the severity level of depression of the participant and the PHQ-8 binary defines whether the participant is depressed or not. For technical reasons, only 182 audio recordings are used. The average length of the recordings is 15 min with a fixed sampling rate of 16 kHz.

5.2 Experimental Setup

Spectrogram-based CNN Architecture: The number of channels of the convolutional and pooling layers are both 128. While their filter size is 3×3 . RELU is used as activation function for all the layers. The stride of the max pooling is 8. The dropout fraction is 0.1.

Table 3. RECOLA dataset results for prediction of arousal. The results obtained for the development and the test sets in term of three metrics: the accuracy, the Pearson’s Coefficient Correlation (PCC) and the Root Mean Square error (RMSE).

	Development			Test		
	Accuracy	PCC	RMSE	Accuracy	PCC	RMSE
MFCC-based CNN	81.93%	0.8130	0.1501	70.23%	0.6981	0.2065
Spectrogram-based CNN	80.20%	0.8157	0.1314	75.65%	0.7673	0.2099
EmoAudioNet	94.49%	0.9521	0.0082	89.30%	0.9069	0.1229

Table 4. RECOLA dataset results for prediction of valence. The results obtained for the development and the test sets in term of three metrics: the accuracy, the Pearson’s Coefficient Correlation (PCC) and the Root Mean Square error (RMSE).

	Development			Test		
	Accuracy	PCC	RMSE	Accuracy	PCC	RMSE
MFCC-based CNN	83.37%	0.8289	0.1405	71.12%	0.6965	0.2082
Spectrogram-based CNN	78.32%	0.7984	0.1446	73.81%	0.7598	0.2132
EmoAudioNet	95.42%	0.9568	0.0625	91.44%	0.9221	0.1118

MFCC-based CNN Architecture: The input is one-dimensional and of size 177×1 . The filter size of its two convolutional layers is 5×1 . RELU is used as activation function for all the layers. The dropout fraction is 0.1 and the stride of the max pooling is 8.

EmoAudioNet Architecture: The two features vectors are concatenated and fed to a fully connected layer of n neurones activated with a Softmax function. n is defined according to the task. When the task concerns binary depression classification, $n = 2$. When the task concerns depression severity level assessment,

		Actual		
		Non-Depression	Depression	Precision
Predicted	Non-Depression	1441 60.52%	354 14.87%	1795 80.28% 19.72%
	Depression	283 11.89%	303 12.73%	586 51.71% 48.29%
Recall		1724 83.58% 16.42%	657 46.12% 53.88%	2381 73.25% 26.75%

Fig. 2. Confusion Matrix of EmoAudioNet generated on the DAIC-WOZ test set

$n = 24$. While, $n = 10$ for arousal or valence prediction. The ADAM optimizer is used. The learning rate is set experimentally to $10e-5$ and it reduced when the loss value stops decreasing. The batch size is fixed to 100 samples. The number of epochs for training is set to 500. An early stopping is performed when the accuracy stops improving after 10 epochs.

5.3 Experimental Results on Spontaneous and Continuous Emotion Recognition from Speech

Results of Three Proposed CNN Architectures. The experimental results of the three proposed architectures on predicting arousal and valence are given in Table 3 and Table 4. EmoAudioNet outperforms MFCC-based CNN and the spectrogram-based CNN with an accuracy of 89% and 91% for predicting arousal and valence respectively. The accuracy of the MFCC-based CNN is around 70% and 71% for arousal and valence respectively. The spectrogram-based CNN is slightly better than the MFCC-based CNN and its accuracy is 76% for predicting arousal and 74% for predicting valence.

EmoAudioNet has a Pearson Coefficient Correlation (PCC) of 0.91 for predicting arousal and 0.92 for predicting valence, and has also a Root Mean Square of Error (RMSE) of 0.12 for arousal’s prediction and 0.11 for valence’s prediction.

Comparisons of EmoAudioNet and the Stat-of-the Art Methods for Arousal and Valence Prediction on RECOLA Dataset. As shown in Table 5, EmoAudioNet model has the best PCC of 0.9069 for arousal prediction. In term of the RMSE, the approach proposed by He *et al.* [12] outperforms all the existing methods with a RMSE equal to 0.099 in predicting arousal.

For valence prediction, EmoAudioNet outperforms state-of-the-art in predicting valence with a PCC of 0.9221 without any fine-tuning. While the proposed approach by He *et al.* [12] has the best RMSE of 0.104.

Table 5. Comparisons of EmoAudioNet and the state-of-the art methods for arousal and valence prediction on RECOLA dataset.

Method	Arousal		Valence	
	PCC	RMSE	PCC	RMSE
He <i>et al.</i> [12]	0.836	0.099	0.529	0.104
Ringeval <i>et al.</i> [11]	0.322	0.173	0.144	0.127
EmoAudioNet	0.9069	0.1229	0.9221	0.1118

5.4 Experimental Results on Automatic Clinical Depression Recognition and Assessment

EmoAudioNet framework is evaluated on two tasks on the DAIC-WOZ corpus. The first task is to predict depression from speech under the PHQ-8 binary. The second task is to predict the depression severity levels under the PHQ-8 scores.

EmoAudioNet Performances on Depression Recognition Task. EmoAudioNet is trained to predict the PHQ-8 binary (0 for non-depression and 1 for depression). The performances are summarized in Fig. 2. The overall accuracy achieved in predicting depression reaches 73.25% with an RMSE of 0.467. On the test set, 60.52% of the samples are correctly labeled with non-depression, whereas, only 12.73% are correctly diagnosed with depression. The low rate of correct classification of non-depression can be explained by the imbalance of the input data on the DAIC-WOZ dataset and the small amount of the participants labeled as depressed. F1 score is designed to deal with the non-uniform distribution of class labels by giving a weighted average of precision and recall. The non-depression F1 score reaches 82% while the depression F1 score reaches 49%. Almost half of the samples predicted with depression are correctly classified with a precision of 51.71%. The number of non-depression samples is twice the number of samples labeled with depression. Thus, adding more samples of depressed participants would significantly increase the model’s ability to recognize depression.

EmoAudioNet Performances on Depression Severity Levels Prediction Task. The depression severity levels are assessed by the PHQ-8 scores ranging from 0 for non-depression to 23 for severe depression. The RMSE achieved when predicting the PHQ-8 scores is 2.6 times better than the one achieved with the depression recognition task. The test loss reaches 0.18 compared to a 0.1 RMSE on the training set.

Comparisons of EmoAudioNet and the State-of-the Art Methods for Depression Prediction on DAIC-WOZ Dataset. Table 6 compares the performances of EmoAudioNet with the state-of-the-art approaches evaluated on the DAIC-WOZ dataset. To the best of our knowledge, in the literature, the best performing approach is the proposed approach in [25] with an F1 score of 85.44%

Table 6. Comparisons of EmoAudioNet and the stat-of-the art methods for prediction of depression on DAIC-WOZ dataset. (*) The results of the depression severity level prediction task. (**) for non-depression. (‡) for depression. (Norm): Normalized RMSE

Method	Accuracy	RMSE	F1 Score
Yang <i>et al.</i> [8]	–	1.46 (*) (depressed male)	–
Yang <i>et al.</i> [26]	–	5.59 (*) (male)	–
Valstar <i>et al.</i> [3]	–	7.78 (*)	–
Al Hanai <i>et al.</i> [23]	–	10.03	–
Salekin <i>et al.</i> [25]	96.7%	–	85.44%
Ma <i>et al.</i> [34]	–	–	70% (**) 50% (‡)
Rejaibi <i>et al.</i> [35]	76.27%	0.4	85% (**) 46% (‡)
	-	0.168 Norm (*)	-
EmoAudioNet	73.25%	0.467	82% (**) 49% (‡)
	-	0.18 Norm 4.14 (*)	-

and an accuracy of 96.7%. The proposed NN2Vec features with BLSTM-MIL classifier achieves this good performance thanks to the leave-one-speaker out cross-validation approach. Comparing to the other proposed approaches where a simple train-test split is performed, giving the model the opportunity to train on multiple train-test splits increase the model performances especially in small datasets.

In the depression recognition task, the EmoAudioNet outperforms the proposed architecture in [34] based on a Convolutional Neural Network followed by a Long Short-Term Memory network. The non-depression F1 score achieved with EmoAudioNet is better than the latter by 13% with the exact same depression F1 score (50%).

Moreover, the EmoAudioNet outperforms the LSTM network in [35] in correctly classifying samples of depression. The depression F1 score achieved with EmoAudioNet is higher than the MFCC-based RNN by 4%. Meanwhile, the overall accuracy and loss achieved by the latter are better than EmoAudioNet by 2.14% and 0.07 respectively. According to the summarized results of previous works in Table 6, the best results achieved so far in the depression severity level prediction task are obtained in [35]. The best normalized RMSE is achieved with the LSTM network to reach 0.168. EmoAudioNet reaches almost the same loss with a very low difference of 0.012. Our proposed architecture outperforms the rest of the results in the literature with the lowest normalized RMSE of 0.18 in predicting depression severity levels (PHQ-8 scores) on the DAIC-WOZ dataset.

6 Conclusion and Future Work

In this paper, we proposed a new emotion and affect recognition methods from speech based on deep neural networks called EmoAudioNet. The proposed EmoAudioNet deep neural networks architecture is the aggregation of an MFCC-based CNN and a spectrogram-based CNN, which studies the time-frequency representation and the visual representation of the spectrum of frequencies of the audio signal. EmoAudioNet gives promising results and it approaches or outperforms state-of-art approaches of continuous dimensional affect recognition and automatic depression recognition from speech on RECOLA and DAIC-WOZ databases. In future work, we are planning (1) to improve the EmoAudioNet architecture with the given possible improvements in the discussion section and (2) to use EmoAudioNet architecture to develop a computer-assisted application for patient monitoring for mood disorders.

References

1. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators: Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015, *Lancet*, vol. 388, no. 10053, pp. 1545–1602 (2015)
2. The National Institute of Mental Health: Depression. <https://www.nimh.nih.gov/health/topics/depression/index.shtml>. Accessed 17 June 2019
3. Valstar, M., et al.: AVEC 2016 - depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Workshop on Audio/visual Emotion Challenge, pp. 3–10. ACM (2016)
4. Ringeval, F., et al.: AVEC 2017 - real-life depression, and affect recognition workshop and challenge. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, pp. 3–9. ACM (2017)
5. Jiang, H., Hu, B., Liu, Z., Wang, G., Zhang, L., Li, X., Kang, H.: Detecting depression using an ensemble logistic regression model based on multiple speech features. *Comput. Math. Methods Medicine* **2018** (2018)
6. Alghowinem, S., et al.: A comparative study of different classifiers for detecting depression from spontaneous speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8022–8026 (2013)
7. Valstar, M., et al.: AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, pp. 3–10 (2013)
8. Yang, L., Sahli, H., Xia, X., Pei, E., Oveneke, M.C., Jiang, D.: Hybrid depression classification and estimation from audio video and text information. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, pp. 45–51. ACM (2017)
9. Cummins, N., Epps, J., Breakspear M., Goecke, R.: An investigation of depressed speech detection: features and normalization. In: Twelfth Annual Conference of the International Speech Communication Association (2011)
10. Lopez-Otero, P., Dacia-Fernandez, L., Garcia-Mateo, C.: A study of acoustic features for depression detection. In: 2nd International Workshop on Biometrics and Forensics, pp. 1–6. IEEE (2014)

11. Ringeval, F., et al.: 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)
12. He, L., Jiang, D., Yang, L., Pei, E., Wu, P., Sahli, H.: Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, pp. 73–80. ACM (2015)
13. Ringeval, F., et al.: AVEC 2018 workshop and challenge: bipolar disorder and cross-cultural affect recognition. In: Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, pp. 3–13. ACM (2018)
14. Dhall, A., Ramana Murthy, O.V., Goecke, R., Joshi, J., Gedeon, T.: Video and image based emotion recognition challenges in the wild: EmotiW 2015. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 423–426 (2015)
15. Haq, S., Jackson, P.J., Edge, J.: Speaker-dependent audio-visual emotion recognition. In: AVSP, pp. 53–58 (2009)
16. Low, L.S.A., Maddage, N.C., Lech, M., Sheeber, L.B., Allen, N.B.: Detection of clinical depression in adolescents' speech during family interactions. *IEEE Trans. Biomed. Eng.* **58**(3), 574–586 (2010)
17. Valstar, M., Schuller, B.W., Krajewski, J., Cowie, R., Pantic, M.: AVEC 2014: the 4th international audio/visual emotion challenge and workshop. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 1243–1244 (2014)
18. Meng, H., Huang, D., Wang, H., Yang, H., Ai-Shuraifi, M., Wang, Y.: Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, pp. 21–30 (2013)
19. Trigeorgis, G., et al.: Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200–5204 (2016)
20. Ringeval, F., et al.: Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recogn. Lett.* **66**, 22–30 (2015)
21. Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Pantic, M.: AVEC 2015: the 5th international audio/visual emotion challenge and workshop. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1335–1336 (2015)
22. Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Topics Signal Process.* **11**(8), 1301–1309 (2017)
23. Al Hanai, T., Ghassemi, M.M., Glass, J.R.: Detecting depression with audio/text sequence modeling of interviews. In: Interspeech, pp. 1716–1720 (2018)
24. Dham, S., Sharma, A., Dhall, A.: Depression scale recognition from audio, visual and text analysis. arXiv preprint [arXiv:1709.05865](https://arxiv.org/abs/1709.05865)
25. Salekin, A., Eberle, J.W., Glenn, J.J., Teachman, B.A., Stankovic, J.A.: A weakly supervised learning framework for detecting social anxiety and depression. *Proc. ACM Interact. Mobile Wearable Ubiquit. Technol.* **2**(2), 81 (2018)
26. Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M.C., Sahli, H.: Multimodal measurement of depression using deep learning models. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, pp. 53–59 (2017)
27. Jain, R.: Improving performance and inference on audio classification tasks using capsule networks. arXiv preprint [arXiv:1902.05069](https://arxiv.org/abs/1902.05069) (2019)

28. Chao, L., Tao, J., Yang, M., Li, Y.: Multi task sequence learning for depression scale prediction from video. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 526–531. IEEE (2015)
29. Gupta, R., Sahu, S., Espy-Wilson, C.Y., Narayanan, S.S.: An affect prediction approach through depression severity parameter incorporation in neural networks. In: Interspeech, pp. 3122–3126 (2017)
30. Kang, Y., Jiang, X., Yin, Y., Shang, Y., Zhou, X.: Deep transformation learning for depression diagnosis from facial images. In: Zhou, J., et al. (eds.) CCBR 2017. LNCS, vol. 10568, pp. 13–22. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69923-3_2
31. Yu, G., Slotine, J.J.: Audio classification from time-frequency texture. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1677–1680 (2009)
32. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8. IEEE (2013)
33. Gratch, J., et al.: The distress analysis interview corpus of human and computer interviews. LREC, pp. 3123–3128 (2014)
34. Ma, X., Yang, H., Chen, Q., Huang, D., Wang, Y.: Depaudionet: an efficient deep model for audio based depression classification. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pp. 35–42 (2016)
35. Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., Othmani, A.: MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. arXiv preprint [arXiv:1909.07208](https://arxiv.org/abs/1909.07208) (2019)
36. Tzirakis, P., Zhang, J., Schuller, B.W.: 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 (2018)