

# EFFICIENT ARABIC EMOTION RECOGNITION USING DEEP NEURAL NETWORKS

Yasser Hifny\*, Ahmed Ali†

\*University of Helwan, Egypt

†Qatar Computing Research Institute, HBKU, Doha, Qatar

yhifny@fci.helwan.edu.eg

amali@qf.org.qa

## ABSTRACT

Emotion recognition from speech signal based on deep learning is an active research area. Convolutional neural networks (CNNs) may be the dominant method in this area. In this paper, we implement two neural architectures to address this problem. The first architecture is an attention-based CNN-LSTM-DNN model. In this novel architecture, the convolutional layers extract salient features and the bi-directional long short-term memory (BLSTM) layers handle the sequential phenomena of the speech signal. This is followed by an attention layer, which extracts a summary vector that is fed to the fully connected dense layer (DNN), which finally connects to a softmax output layer. The second architecture is based on a deep CNN model. The results on an Arabic speech emotion recognition task show that our innovative approach can lead to significant improvements (2.2% absolute improvements) over a strong deep CNN baseline system. On the other hand, the deep CNN models are significantly faster than the attention based CNN-LSTM-DNN models in training and classification.

**Index Terms**— Speech emotion recognition, deep neural networks (DNN), convolutional neural networks (CNN), bi-direction long short-term memory (BLSTM) networks, and attention.

## 1. INTRODUCTION

Deep architectures for emotion recognition from speech is a growing research field [1, 2, 3, 4, 5]. Using short time signal analysis, a speech utterance is represented by a matrix  $X \in R^{d \times T}$  where  $T$  is the size of the time dimension and  $d$  is the size of spectral dimension. The sequence to sequence layers model the spectral phenomena and keep the size of the time dimension  $T$  without any modification. A sequence to vector layer is used to convert the sequence to a fixed dimension vector which can be fed to feed forward dense layers. The global average pooling, global max pooling and attention are common choices for this type of layer. Fully-connected dense layers are then used to apply nonlinear compression of the input features for better representation which improves the modeling power of the classifier. A multiclass classifier

is implemented using a softmax layer. Typically, the model is trained using the cross-entropy objective function.

Convolutional neural networks (CNNs) have been recently used in many emotion recognition tasks. For example, CNNs designed for visual recognition (e.g. AlexNet) were directly adapted for emotion recognition from speech [5]. Moreover, in a study by [3], they conducted extensive experiments using an attentive CNN with multi-view learning objective function using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [6]. They concluded that for a CNN architecture, the particular choice of features is not as important as the model architecture, or the amount and kind of training data. CNNs are excellent in feature extraction and very fast in training compared to standard sequence modeling.

Long short-term memory networks (LSTMs) [7] sequence to sequence layers are excellent in capturing the sequential phenomena of the speech signal for various style of speaking. In a study by [8], they propose a solution to the problem of ‘context-aware’ emotional relevant feature extraction, by combining CNNs with LSTM networks, in order to automatically learn the best representation of the speech signal directly from the raw time representation. They did not use any of the commonly hand-engineered features, such as mel-Frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) coefficients. Their end-to-end system was targeted to learn an intermediate representation of the raw input signal automatically that better suits the task at hand and hence leads to better performance.

Both CNN and LSTM networks have shown significant improvements over fully-connected neural network across a wide variety of tasks. In recent work by [9], they took advantage of CNNs, LSTMs and DNNs by combining them into one unified architecture for speech recognition task. CNNs are good at reducing frequency variations, LSTMs are good at temporal modeling, and finally DNNs map the features into a more separable space. Their CLDNN provided a 4-6% relative improvement in WER. In a similar work for emotion recognition from speech [10], the combination of CNNs and LSTMs led to improvements in the classification accuracy. The last state of the LSTM was used for sequence to vector conversion.

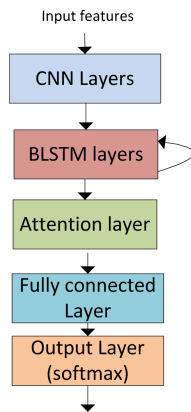
Recently, an end-to-end multimodal emotion and gender recognition model with dynamic joint loss weights is developed [11]. The proposed model does not need any pre-trained features from audio and visual data. In addition, the system is trained using a multitask objective function and its weights are assigned using a dynamic approach.

In this paper, we build on these contributions to develop an emotion recognition system for Arabic data using the recently introduced KSU emotions corpus<sup>1</sup>. Our contributions are: (i) Introducing a novel approach for emotion recognition by using an attention based CNN-LSTM-DNN architecture; (ii) Studying a deep CNN models for the same task; (iii) Comparing our results with published state-of-the-art results on the IEMOCAP database and (iii) Providing our scripts and code for the research community for usage and potential future contributions<sup>2</sup>.

The rest of the paper is organized as follows: In section 2, we describe the attention-based CNN-LSTM-DNN and the proposed deep CNN architectures. Data is explained in section 3. Experimental setup is illustrated in section 4. This is followed by results in section 5. Finally section 6 concludes the paper and discusses future work.

## 2. PROPOSED APPROACH

Our main emotion classifier is based on the CNN-LSTM-DNN architecture as shown in Figure 1. The CNN layers [12] extract features from the audio signal, while the bi-directional LSTM (BLSTM) layers [7, 13] handle the sequential phenomena of the speech signal. This is followed by an attention layer, which extracts a summary vector that is fed to a DNN layer, which finally connects to a softmax layer.



**Fig. 1.** Attention based CNN-BLSTM-DNN architecture for speech emotion recognition.

<sup>1</sup><https://catalog.ldc.upenn.edu/LDC97S45>

<sup>2</sup><http://github.com/qcri/deepemotion>

## 2.1. Attention layer

The output of the BLSTM layer is followed by an attention layer, which extracts a summary vector that is fed to the fully connected dense layer (DNN), which finally connects to a softmax layer [3]. For each vector  $\mathbf{x}_t$  in a sequence of inputs  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$ , the attention weights  $\alpha_t$  are given by:

$$\alpha_t = \frac{\exp(f(\mathbf{x}_t))}{\sum_{j=1}^T \exp(f(\mathbf{x}_j))} \quad (1)$$

where  $f(\mathbf{x}_t)$  is defined using the trainable parameters  $\mathbf{w}$  as follows:

$$f(\mathbf{x}_t) = \tanh(\mathbf{w}^T \mathbf{x}_t) \quad (2)$$

The summary vector which is output of the attention layer is computed as a weighted sum of the input sequence:

$$C = \sum_{t=1}^T \alpha_t \mathbf{x}_t \quad (3)$$

Since most of the research in this area is based on CNN models [3, 5], we have implemented a strong CNN baseline.

## 3. DATA

We use the newly developed Arabic speech emotion database (KSUEmotions) [14] for our experiments. The corpus was designed for Modern Standard Arabic (MSA) and was recorded using 23 speakers (10 males and 13 females). The speakers are from three Arab countries: Yemen, Saudi Arabia and Syria. The speech was recorded in two phases. In Phase-1, twenty speakers were selected (ten male speakers from Saudi Arabia, Yemen, and Syria and ten female speakers from two countries Saudi Arabia and Syria). In this phase, neutral, sadness, happiness, surprise, and questioning emotions were selected.

Seven male speakers from phase one and seven female speakers (four of them from phase one and the other new three female speakers were added from Yemen nationality) were selected in phase two. In this phase, the questioning emotion was excluded while the anger emotion was added to this corpus to be more consistent with other similar corpora. The total duration of the all recorded files was 2 hours and 55 minutes for Phase-1 and 2 hours and 15 minutes for Phase-2. We mixed Phase-1 and Phase-2 datasets to generate a bigger corpus (Phase-mix). Table 1 summarizes the statistics of this corpus.

In an attempt to validate our findings, we built a strong baseline model using four emotional categories (neutral, angry, sad and happy) from the interactive emotional dyadic motion capture (IEMOCAP) database [6]. The database consists of about 12 hours of audiovisual data (speech, video, facial motion capture) from five mixed gender pairs of male and female actors, at two recording scenarios: scripted and improvised speech. The IEMOCAP corpus have five sessions,

**Table 1.** KSU emotions dataset statistics.

Phase	Duration	Supported emotions
Phase1	3.0 hrs	neutral, sadness, happiness, surprise, and questioning
Phase2	2.2 hrs	neutral, sadness, happiness, surprise, and anger
Mixed	5.2 hrs	neutral, sadness, happiness, surprise, questioning, and anger

four of which are used for training and the remaining one is used for testing. Each wave file is manually labeled and has segment level emotion category.

## 4. EXPERIMENTAL SETUP

### 4.1. Input features

We use the Essentia toolkit [15] to extract 13 MFCCs. The audio signal is segmented into 25 ms long frames with a 10 ms shift. A Hamming window is applied and the FFT with 512 points is computed. Then, we compute the logarithmic power of 26 Mel-frequency filter-banks over a range from 0 to 8 kHz. Finally, a discrete cosine transform (DCT) is applied to extract the first 13 MFCCs. These features are used to train our emotion classifiers.

### 4.2. Training platform

We implemented the emotion classification models using Keras deep learning library version 2.0 with Tensorflow 1.6.0 [16] backend. The models were trained using an Nvidia's Tesla K80 graphical processing unit (GPU). We used the Adam optimization algorithm [17] to train our models with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . The initial learning rate was set to 0.001 and the batch size was set to 32. Shorter sequences of a batch were padded with zeros to match the longest sequence in the data (i.e. 15.5 seconds).

### 4.3. Models

A strong deep CNN emotion classifier was developed as a baseline. It has six layers where four of them are CNN layers (see Table 2). Dropout rate between layers was set to 0.2. We benchmarked our results with the well-known IEMOCAP database [6].

Our experiment led to classification accuracy of 56.0% which is similar to the state-of-the art results on this task [3]. Hence, we used this model as a baseline to train the Arabic emotion recognition task. Our CNN-LSTM-DNN emotion classifier has nine layers (see Table 3). Both models are trained using the crossentropy objective function.

**Table 2.** The deep CNN architecture.

Layer	Type	Details
1	Conv	500 filters + Relu + Stride=1 + kernel width=5
2	Conv	500 filters + Relu + Stride=2 + kernel width=7
3	Conv	500 filters + Relu + Stride=2 + kernel width=1
4	Conv	500 filters + Relu + Stride=1 + kernel width=1
5	MaxPool1D	Global
6	Softmax	6 classes for Phase-mix dataset

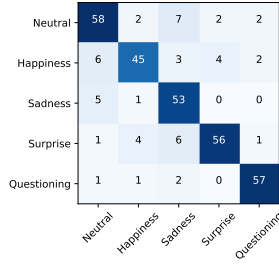
**Table 3.** The CNN-LSTM-DNN architecture.

Layer	Type	Details
1	Conv	256 filters + Relu + Stride=1 + kernel width=5
2	MaxPool1D	Pool size=2
3	Conv	64 filters + Relu + Stride=1 + kernel width=5
4	MaxPool1D	Pool size = 2
5	BLSTM	128 units (64 in each direction)
6	BLSTM	128 units (64 in each direction)
7	Attention	128 dimension
8	Dense	64 units + tanh activation
9	Softmax	6 classes for Phase-mix dataset

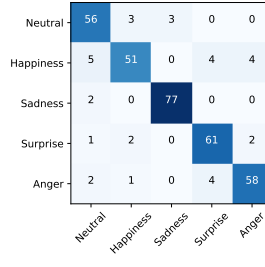
## 5. RESULTS

Our results were generated using Phase-1, Phase-2 and Phase-mix datasets. We ran five folds cross validation experiments, the results over the five folds were consistent. Hence, we report the results for fold number five experiments for each phase. In Figure 2, the confusion matrices for each experiment are detailed. We observe that in Phase-1 experiment, the questioning and sadness emotions were recognized successfully most of the time. In Phase-2, sadness emotion was confused less than other emotions. In Phase-mix, the classification problem seems to be harder than Phase-1 and Phase-2 datasets. The overall precision, recall and F1 scores for each emotion are shown in Figure 3. It is worth noting that both emotions, questioning and anger performed poorly compared to the rest of the emotions in the Phase-mix as shown in Figure 5. This is primarily due to less data being used for these two emotions in the combined dataset.

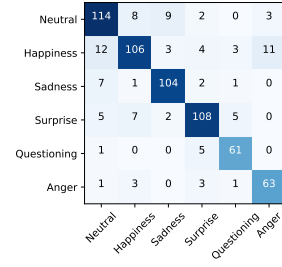
Table 4 shows the results for both the attention-based CNN-BLSTM-DNN and deep CNN systems. The attention-



(a) Confusion matrix for Phase-1.

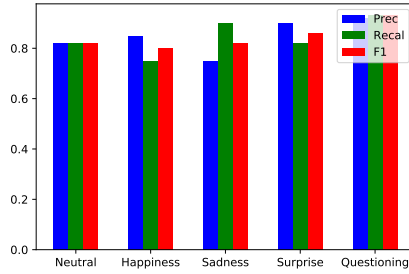


(b) Confusion matrix for Phase-2.



(c) Confusion matrix for Phase-mix.

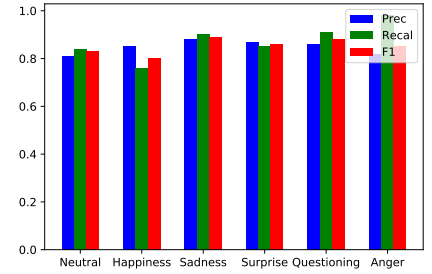
**Fig. 2.** Confusion matrix for different phases. Note that the  $x$  axis shows predictions and the  $y$  axis shows the labels.



(a) The F1 results for Phase-1 dataset.



(b) The F1 results for Phase-2 dataset.



(c) The F1 results for Phase-mix dataset.

**Fig. 3.** Overall F1 results for all the different emotion classes.

based CNN-BLSTM-DNN model results are consistently better than the deep CNN model. When the attention layer is replaced with last state of the second LSTM layer as in [10], the classification accuracy of the model drops significantly (i.e. more than 2% absolute). On the other hand, the CNN baseline is much faster than the attention-based CNN-BLSTM-DNN models in training and classification.

**Table 4.** Classification results using the Phase-mix dataset.

Fold	Classification overall accuracy in %	
	CNN-BLSTM-DNN	CNN
1	87.7	85.2
2	86.6	85.6
3	88.2	84.4
4	86.0	85.2
5	87.6	84.7
Average	87.2	85.0

## 6. CONCLUSIONS

In this paper, we designed an end-to-end attention-based CNN-LSTM-DNN emotion classifier. In our classifier, the convolutional layers (CNN) extract salient features, the bi-directional long short-term memory (BLSTM) layers handle the sequential phenomena of the speech signal. This is followed by an attention layer, which extracts a summary vector that is fed to the fully connected dense layer (DNN), which finally connects to a softmax layer. The results on an Arabic speech emotion recognition task show that our innovative approach can lead to significant improvements (2.2% absolute improvements) over a strong deep CNN baseline system. However, the deep CNN models are significantly faster than the attention-based CNN-LSTM-DNN models in training and classification phases.

Future work will focus on training an ensemble classifier and interpolating the predictions to improve the classification accuracy. We plan to use large Arabic emotion databases using our powerful attention-based CNN-LSTM-DNN models. In addition, joint estimation of the emotion, dialect, language, and accent using multitask learning will be investigated. In addition, the separate label per frame methods developed in [4] will be compared with the single label per utterance methods commonly used in the field in a unified framework.

## 7. REFERENCES

- [1] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos, “Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011,” *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.
- [2] Björn W Schuller, “Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends,” *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.
- [3] 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.
- [4] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak, “Emotion identification from raw speech signals using dnns,” *Proc. Interspeech 2018*, pp. 3097–3101, 2018.
- [5] Yuanyuan Zhang, Jun Du, Zirui Wang, and Jianshu Zhang, “Attention based fully convolutional network for speech emotion recognition,” *arXiv preprint arXiv:1806.01506*, 2018.
- [6] 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*, vol. 42, no. 4, pp. 335, 2008.
- [7] Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [8] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalís 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*. IEEE, 2016, pp. 5200–5204.
- [9] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4580–4584.
- [10] Aharon Satt, Shai Rozenberg, and Ron Hoory, “Efficient emotion recognition from speech using deep learning on spectrograms,” *Proc. Interspeech 2017*, pp. 1089–1093, 2017.
- [11] Myungsu Chae, Tae-Ho Kim, Young Hoon Shin, Jun-Woo Kim, and Soo-Young Lee, “End-to-end multi-modal emotion and gender recognition with dynamic weights of joint loss,” *arXiv preprint arXiv:1809.00758*, 2018.
- [12] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn, “Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4277–4280.
- [13] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber, “Bidirectional lstm networks for improved phoneme classification and recognition,” in *International Conference on Artificial Neural Networks*. Springer, 2005, pp. 799–804.
- [14] Ali Meftah, Yousef A Alotaibi, and Sid-Ahmed Selouani, “Designing, building, and analyzing an arabic speech emotional corpus: Phase 2,” in *5th International Conference on Arabic Language Processing*, 2014, pp. 181–184.
- [15] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Perfecto Herrera Boyer, Oscar Mayor, Gerard Roma Trepas, Justin Salamon, José Ricardo Zapata González, and Xavier Serra, “Essentia: An audio analysis library for music information retrieval,” in *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8. International Society for Music Information Retrieval (ISMIR), 2013.*
- [16] François Chollet et al., “Keras,” 2015.
- [17] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.