ASERS-CNN: ARABIC SPEECH EMOTION RECOGNITION SYSTEM BASED ON CNN MODEL

Mohammed Tajalsir[1], Susana Mu˜noz Hern´andez[2]
and Fatima Abdalbagi Mohammed[3]

[1]Department of Computer Science,
Sudan University of Science and Technology, Khartoum, Sudan
[2]Technical University of Madrid (UPM),
Computer Science School (FI), Madrid, Spain
[3]Department of Computer Science,
Sudan University of Science and Technology, Khartoum, Sudan

ABSTRACT

When two people are on the phone, although they cannot observe the other person's facial expression and physiological state, it is possible to estimate the speaker's emotional state by voice roughly. In medical care, if the emotional state of a patient, especially a patient with an expression disorder, can be known, different care measures can be made according to the patient's mood to increase the amount of care. The system that capable for recognize the emotional states of human being from his speech is known as Speech emotion recognition system (SER). Deep learning is one of most technique that has been widely used in emotion recognition studies, in this paper we implement CNN model for Arabic speech emotion recognition. We propose ASERS-CNN model for Arabic Speech Emotion Recognition based on CNN model. We evaluated our model using Arabic speech dataset named Basic Arabic Expressive Speech corpus (BAES-DB). In addition of that we compare the accuracy between our previous ASERS-LSTM and new ASERS-CNN model proposed in this paper and we comes out that our new proposed mode is outperformed ASERS-LSTM model where it get 98.18% accuracy.

KEYWORDS:

BAES-DB, ASERS-LSTM, Deep learning, Speech emotion recognition.

1. INTRODUCTION

Speech is the compound words that are useful in the situation and used for communication, to express the thoughts and opinions. The emotions are the mental states such as happiness, love, fear, anger, or joy that effect on the human behavior. is easy for the Human using available senses to detect the emotional states from speaker's speech , but this is a very difficult task for machines. The research work on speech emotion recognition generally starts with the study of phonetic features in the field of phonetics, and some studies make use of the characteristics of linguistics. For example, some special vocabulary, syntax, etc., in general, all studies have shown that the performance of automatic speech emotion recognition is inferior to the ability of humans to recognize emotions.

The Arabic language is spoken by more than 300 million Arabic speakers. Since that, there are needs for developing an Arabic Speech Recognition system to make the computer not just can recognize the speech, but to recognize how has spoken. However, the Arabic Speech Recognition system is challenged by many factors such as the cultural effects and determining the suitable

DOI: 10.5121/sipij.2022.13104

features that classify the emotions. There are few Arabic speech dataset to be used for training Arabic Speech Recognition models, the most difficult task related the speech samples is to find reliable data, most of dataset are not public, and most of datasets related sentiment analysis are not real or simulated. And it is very difficult to record people's real emotions. In some works related to detecting lying or stress you can force the people to lie or be stressed. But this is very difficult in other interesting emotions as happiness, sadness, anger.

Deep Learning is a new area of Machine Learning research, the relationship between the artificial intelligence.

There are few research work that was focused on Arabic emotions recognition such as the work in [2], [3], [4], [5], [6], [7] and [8]. In [2] recognizing emotions was performed using Arabic natural real life utterances for the first time. The realistic speech corpus collected from Arabic TV shows and the three emotions happy, angry and surprised are recognized. The low-level descriptors (LLDs); acoustic and spectral features extracted and calculate 15 statistical functions and the delta coefficient is computed for every feature, then ineffective features removed using Kruskal–Wallis non-parametric test leading to new database with 845 features. Thirty-five classifiers belonging to six classification groups Trees, Rules, Bayes, Lazy, Functions and Metawere applied over the extracted features. The Sequential minimal optimization (SMO) classifier (Functions group) overcomes the others giving 95.52% accuracy.

in [3] they enhancement the emotion recognition system by proposed a new two phase, the first phase aim to remove the units that were misclassified by most classifiers from the original corpora by labelling each video unit misclassified by more than twenty four methods as "1" and good units as "0" and phase two remove all videos with type "1" from original database than all classification models over the new database. The new enhancement model improved the accuracy by 3% for all for all thirty five classification models. The result accuracy of Sequential minimal optimization (SMO) classifier improved from 95.52% to 98.04%.

[4] They performed two neural architectures to develop an emotion recognition system for Arabic data using KSUEmotions dataset; an attention-based CNNLSTM-DNN model and a strong deep CNN model as a baseline. In the first emotion classifier the CNN layers used to extract audio signal features and the bi-directional LSTM (BLSTM) layers used to process the sequential phenomena of the speech signal then followed by an attention layer to extracts a summary vector which is fed to a DNN layer which finally connects to a softmax layer. The results show that an attention-based CNNLSTM-DNN approach can produce to significant improvements (2.2% absolute improvements) over the baseline system.

A semi-natural Egyptian Arabic speech emotion (EYASE) database is introduced by [5], the EYASE database has been created from Egyptian TV series. Prosodic, spectral and wavelet features in addition to pitch, intensity, formants, Mel-frequency cepstral coefficients (MFCC), long-term average spectrum (LTAS) and wavelet parameters are extracted to recognize four emotions: angry, happy, neutral and sad. Several experiments were performed to detect emotions: emotion vs. neutral classifications, arousal & valence classifications and multi-emotion classifications for both speaker independent and dependent experiments. The experiments analysis finds that the gender and culture effects on SER. Furthermore, For the EYASE database, anger emotion most readily detected while happiness was the most challenging. Arousal (angry/sad) recognition rates were shown to be superior to valence (angry/happy) recognition rates. In most cases the speaker dependent SER performance overcomes the speaker independent SER performance.

In [6] five ensemble models {Bagging, Adaboost, Logitboost, Random Subspace and Random Committee} were employed and studied their effect on a speech emotions recognition system. The highest accuracy was obtained by SMO 95.52% through all single classifiers for recognizing happy, angry, and surprise emotion from The Arabic Natural Audio Dataset (ANAD). And after applying the ensemble models on 19 single classifiers the result accuracy of SMO classifier improved from 95.52% to 95.95% as the best enhanced. The Boosting technique having the Naïve Bayes Multinomial as base classifier achieved the highest improvement in accuracy 19.09%.

In [7] they built a system to automatically recognize emotion in speech using a corpus of Arabic expressive sentences phonetically balanced. They study the influence of the speaker dependency on their result. The joy, sadness, anger and neutral emotions were recognized after extracts the cepstral features, their first and second derivatives, Shimmer, Jitter and the duration. They used a multilayer perceptron neural network (MLPs) to recognize emotion. The experiments shows that in the case of an intra-speaker classification recognition rate could reach more than 98% and in the case of an inter-speaker classification recognition rate could reach 54.75%.thus the system's dependence on speaker is obvious.

A natural Arabic visual-audio dataset was designed by [8] which consist of audio-visual records from the Algerian TV talk show “Red line” for the present. The dataset was records by 14 speakers with 1, 443 complete sentences speech. The openSMILE feature extraction tool used to extracts variety of acoustic features to recognize five emotions enthusiasm, admiration, disapproval, neutral, and joy. The min, max, range, standard deviation and mean statistical functions are applied over all the extracted features (energy, pitch, ZCR, spectral features, MFCCs, and Line Spectral Frequencies (LSP)). The WEKA toolkit was used to perform some classification algorithms. The experiments results shows that using the SMO classifier with the Energy, pitch, ZCR, spectral features, MFCCs, LSP features set (430 features) achieves the better classification results (0.48) which are measured by a weighted average of f-measure. And recognizing the Enthusiasm is the most difficult task through the five emotions.

The rest of this paper is organized as flow: section 2 briefly describes the methods and techniques used in this work. The results are detailed and discussed in section 3. Finally, in section 4 the conclusions are drawn.

2. METHODS AND TECHNIQUES

Pre-processing the signal and Feature Extraction, training the model and finally testing the model these are the phases for our methodology to recognize the Emotion as we can see in Figure 2.

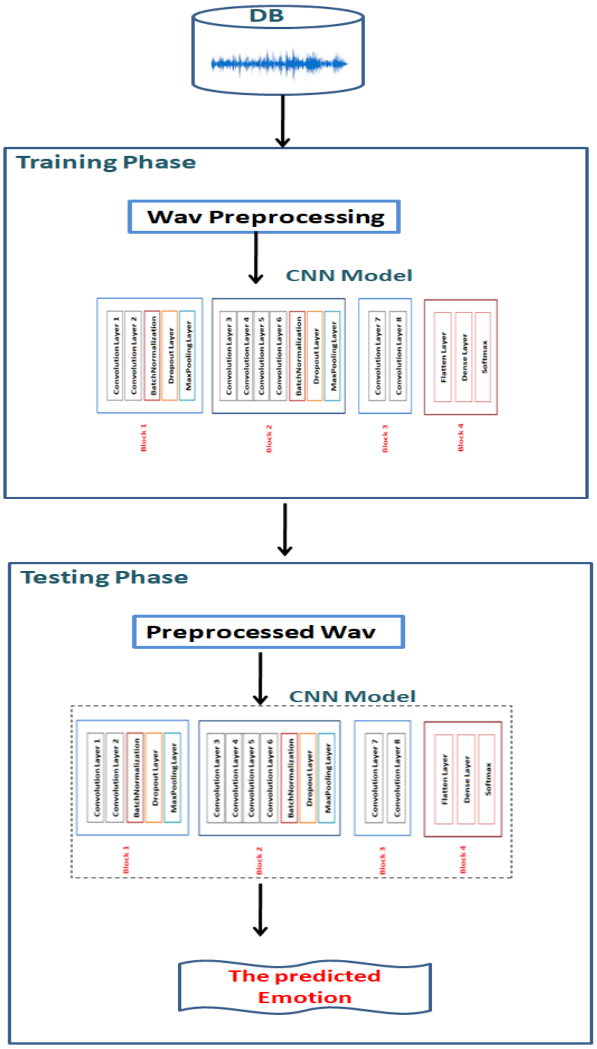


Figure 1: The Methodology to recognize Emotion

2.1. Dataset

Basic Arabic Expressive Speech corpus (BAES-DB) is the dataset that we used for our experimental. The corpus consists of 13 speakers, 4 emotions and 10 sentences; in total it contains 520 sound files. Each file of them contains the ten utterances of a sentence in one of the four emotions states for each of the thirteen speakers. The four selected emotions are neutral, joy, sadness and anger. Any speaker recorded the all ten sentences in a specific situation before moving on to the next one. The first four speakers were recorded while sitting, whereas the 9 others were standing.

2.2. Preprocessing

For preprocessing step, Pre-emphasis and Silent Remover are the two preprocessing algorithms that we apply it in here in this paper same as we did in our previous work.

2.3. Feature Extraction

Human beings can assume the emotional state of the speaker by the voice of the other party. Subjectively, we know that when a person is angry, his speech rate will increase, the loudness of the voice will increase, and the tone will increase. When a person is sad, the speech rate will slow down, the loudness will decrease, the tone will decrease, etc. This shows that the human voice situation will be significantly affected by the speaker's emotions. When people are in different emotions, the sounds produced will have different acoustic characteristics or generally call it as Feature. So we need firstly to extract the feature from speech to detect the speaker's emotions. Here in this paper five different feature types are investigated using the CNN architecture: Mel-Frequency Cepstral Coefficients (MFCC) features, chromagram, Mel-scaled spectrogram, spectral contrast and tonal centroid features (tonnetz).

2.4. Convolutional Neural Network

The convolutional neural network (CNN) defines as: one of the most popular algorithms for deep learning with images and video [10]. CNN it's composed of an input layer, an output layer, and many hidden layers in between just like other neural networks.

2.4.1. Training CNN model

We implemented the CNN model for emotion classification using Keras deep learning library with Tensorflow backend on laptop with processor Intel(R) Core™ i7-3520 CPU @ 2.90GHz, 8GB RAM and 64 bit operating system. The ASERS-CNN model Architecture proposed here in this paper is consist of 4 block, each one of them contain of the convolution, Batch Normalization, Dropout and MaxPooling layer as shown in Figure 4.

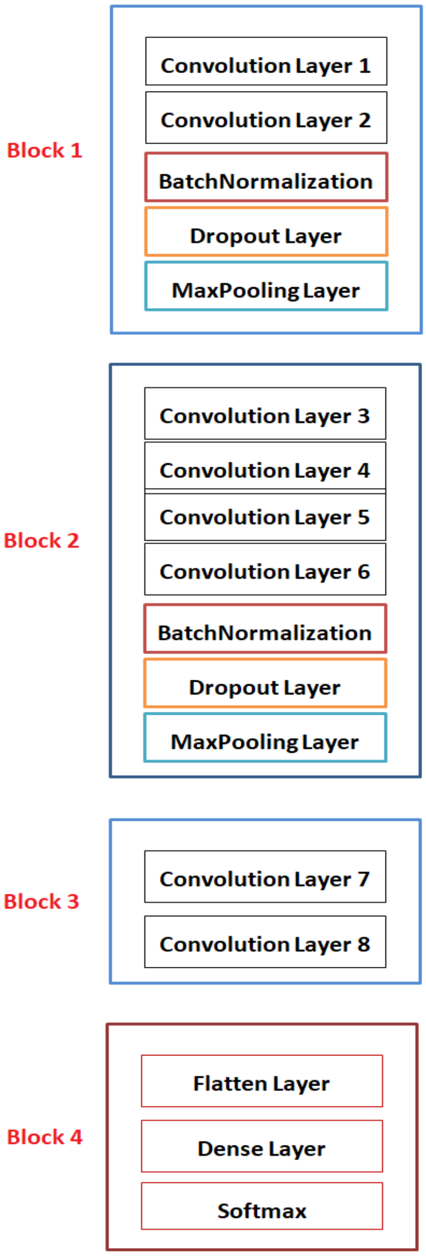
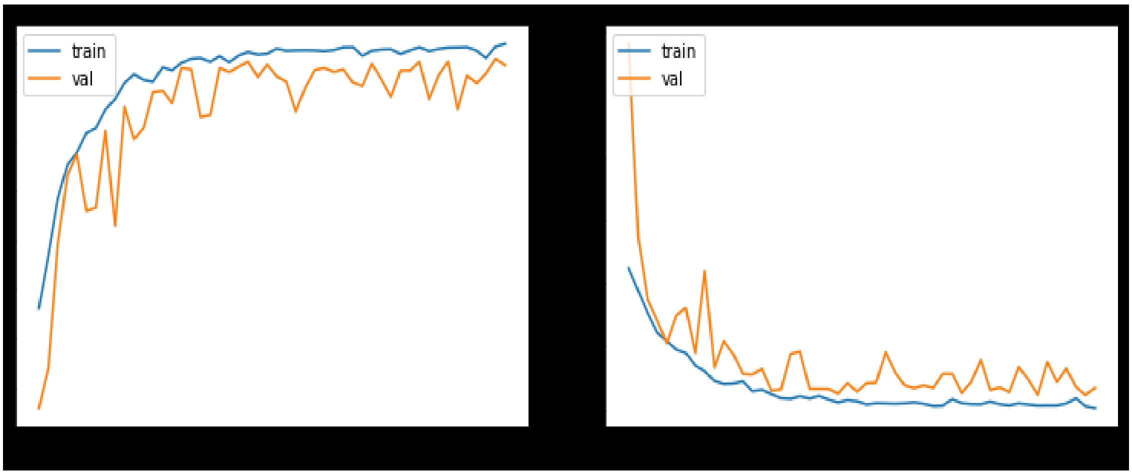


Figure 2: ASERS-CNN model

The model was trained using Adam optimization algorithm with dropout rate=0.2. The initial learning rate was set to 0.001 and the batch size was set to 32 for 50 Epoch, Categorical cross entropy loss function was used. The accuracy and loss curves for the model are shown in Figure 5.



(a) **(b)**

Figure 3: (a) CNN Model Accuracy curve (b) CNN Model loss curve

3. Results and Discussion

In order to evaluate the performance of the proposed CNN model we used Basic Arabic Expressive Speech corpus (BAES-DB) database, which 520 wav files for 13 speakers. For training and evaluation the model, we used four categorical emotions Angry, Happy, Sad and Neutral, which represent the majority of the emotion categories in the database. For low-level acoustic features, we extract 5 features: Mel-Frequency Cepstral Coefficients (MFCC) features, chromagram, Mel-scaled spectrogram, spectral contrast and tonal centroid features (tonnetz). To determine whether the number of Epoch for the training has any effect on the accuracy of emotion classification, we was trained the model using 200 and 50 Epoch. Also we study the effect of the number of feature that extracted from each wav file.

In Table 1 we summarize the Compression between the three models DNN, LSTM and CNN Where the DNN its get Accuracy 93.34% before applying the Pre-processing step and after Pre-processing it get 97.78% when using 5 Feature. But when using 7 Feature and before Pre-processing is getting 92.95% and after Pre-processing it gets 98.06%. When decrease number of Epoch from 200 to 50 it get 97.04%. The CNN model its get Accuracy 98.18% before applying the Pre-processing step and after Pre-processing it get 98.52% when using 5 Feature. But when using 7 Feature and before Pre-processing is getting 98.40% and after Pre-processing it gets 98.46%. When decrease number of Epoch from 200 to 50 it get 95.79%.

The LSTM model its get Accuracy 96.81% before applying the Pre-processing step and after Pre-processing it get 97.44% when using 5 Feature. But when using 7 Feature and before Pre-processing is getting 96.98% and after Pre-processing it gets 95.84%. When decrease number of Epoch from 200 to 50 it get 95.16%. As we can see in Table 1, the results show the best Accuracy for CNN model is when we use 200 epochs with two pre-processing and five extracted features.

Table 1: Compression the Accuracy between CNN, DNN and LSTM models

		The Accuracy (%)				
		200 Epoch				**50 Epoch**
		5 Feature		7 Feature		7 Feature With Preprocessing
		Without Preprocessing	With Preprocessing	Without Preprocessing	With Preprocessing	
Classifier	DNN	**93.34%**	**97.78%**	**92.95%**	**98.06%**	**97.04%**
	CNN	**98.18%**	**98.52%**	**98.40%**	**98.46%**	**95.79%**
	LSTM	**96.81%**	**97.44%**	**96.98%**	**95.84%**	**95.16%**

The results are compared with some of related works include (Klaylat et al., 2018a), (Klaylat et al., 2018b), (Schuller, Rigoll and Lang, 2014),(Shaw, 2016) and (Farooque et al., 2004) as shown in Table 2 bellow. Our Proposed CNN it obviously from the Table 2 it's outperformed the other state-of-the-art model which it's obtain 98.52%. (Klaylat et al., 2018b) it's obtaining better Accuracy than Our Proposed LSTM where it get 98.04% and Our Proposed LSTM it get 97.44%. Our Proposed DNN it's outperformed (Klaylat et al., 2018a) , [12], [13] and [14].

Table 2: Compression between our proposed models and some of related works

Ref	**Classifier**	**Features**	**Database**	**Accuracy**
[2]	Thirty-five classifiers belonging to six groups	low-level descriptors (LLDs); acoustic and spectral features	created database	95.52%
[3]	Thirty-five classifiers belonging to six groups	low-level descriptors (LLDs); acoustic and spectral features	created database	98.04%.
[12]	CHM, GMM.	Pitch and energy.	-	78%
[13]	ANN	Pitch, Energy, MFCC, Formant Frequencies	created database	85%.
[14]	RFuzzy model	SP-IN, TDIFFVUV, and TDIFFW	created database	90%
Our DNN	**DNN Network**	**Mel-Frequency Cepstral Coefficients (MFCC) features, chromagram, Mel-scaled spectrogram, spectral contrast and tonal centroid features**	**BAES-DB**	**98.06%**
Our CNN	**CNN Network**			**98.52%**
Our LSTM	**LSTM Network**			**97.44%**

4. CONCLUSION

in This paper we propose ASERS-CNN Deep learning model for Arabic speech emotion recognition (ASER),we use Arabic speech dataset named Basic Arabic Expressive Speech corpus (BAES-DB) for our evaluation the model get Accuracy 98.18% before applying any Pre-processing and after Pre-processing it get 98.52% when using 5 Feature. When we use 7 Feature and before Pre-processing are getting 98.40% and after Pre-processing it get 98.46%. we also decrease number of Epoch from 200 to 50 and it get 95.79%.We compared the accuracy of proposed ASERS-CNN model with our previous LSTM and DNN models. The proposed ASERS-CNN model it's outperformed the previous LSTM and DNN model where it gets 98.52% accuracy and LSTM and DNN was gets 98.06% and 97.44%, respectively. We also compared the accuracy of proposed ASERS-CNN model with some of related works include (Klaylat et al., 2018a), (Klaylat et al., 2018b), (Schuller, Rigoll and Lang, 2014),(Shaw, 2016) and (Farooque et al., 2004) as shown in Table 2 bellow. Our Proposed CNN it obviously from the Table 2 it's outperformed the other state-of-the-art model.

REFERENCES

[1] F. Chollet, *Deep Learning with Python*, vol. 10, no. 2. 2011.

[2] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, “Emotion recognition in Arabic speech,” *Analog Integr. Circuits Signal Process.*, vol. 96, no. 2, pp. 337–351, 2018.

[3] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, “Enhancement of an Arabic Speech Emotion Recognition System,” *Int. J. Appl. Eng. Res.*, vol. 13, no. 5, pp. 2380–2389, 2018.

[4] Y. Hifny and A. Ali, “Efficient Arabic emotion recognition using deep neural networks,” pp. 6710–6714, 2019.

[5] L. Abdel-Hamid, “Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features,” *Speech Commun.*, vol. 122, pp. 19–30, 2020.

[6] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, “Ensemble Models for Enhancement of an Arabic Speech Emotion Recognition System,” vol. 70, no. January, pp. 299–311, 2020.

[7] I. Hadjadji, L. Falek, L. Demri, and H. Teffahi, “Emotion recognition in Arabic speech,” *2019 Int. Conf. Adv. Electr. Eng. ICAEE 2019*, 2019.

[8] H. Dahmani, H. Hussein, B. Meyer-Sickendiek, and O. Jokisch, “Natural Arabic Language Resources for Emotion Recognition in Algerian Dialect,” vol. 2, no. October, pp. 18–33, 2019.

[9] B. Y. Goodfellow Ian and Courville Aaron, *Deep learning 简介一、什么是 Deep Learning ？*, vol. 29. 2019.

[10] Mathworks, “Introducing Deep Learning with MATLAB,” *Introd. Deep Learn. with MATLAB*, p. 15, 2017.

[11] S. Hung-Il, “Chapter 1 - An Introduction to Neural Networks and Deep Learning,” *Deep Learning for Medical Image Analysis*. pp. 3–24, 2017.

[12] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov Model-based Speech Emotion Recognition,” no. August 2003, 2014.

[13] A. Shaw, “Emotion Recognition and Classification in Speech using Artificial Neural Networks,” vol. 145, no. 8, pp. 5–9, 2016.

[14] M. Farooque, S. Munoz-hernandez, C. De Montegancedo, and B. Monte, “Prototype from Voice Speech Analysis,” 2004.