ASERS-LSTM: ARABIC SPEECH EMOTION RECOGNITION SYSTEM BASED ON LSTM 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

The swift progress in the study field of human-computer interaction (HCI) causes to increase in the interest in systems for Speech emotion recognition (SER). The speech Emotion Recognition System is the system that can identify the emotional states of human beings from their voice. There are well works in Speech Emotion Recognition for different language but few researches have implemented for Arabic SER systems and that because of the shortage of available Arabic speech emotion databases. The most commonly considered languages for SER is English and other European and Asian languages. Several machine learning-based classifiers that have been used by researchers to distinguish emotional classes: SVMs, RFs, and the KNN algorithm, hidden Markov models (HMMs), MLPs and deep learning. In this paper we propose ASERS-LSTM model for Arabic Speech Emotion Recognition based on LSTM model. We extracted five features from the speech: Mel-Frequency Cepstral Coefficients (MFCC) features, chromagram, Mel-scaled spectrogram, spectral contrast and tonal centroid features (tonnetz). We evaluated our model using Arabic speech dataset named Basic Arabic Expressive Speech corpus (BAES-DB). In addition of that we also construct a DNN for classify the Emotion and compare the accuracy between LSTM and DNN model. For DNN the accuracy is 93.34% and for LSTM is 96.81%.

KEYWORDS

Emotion recognition, Deep learning, LSTM, DNN.

1. INTRODUCTION

Voice is the sound of human beings it composed by the succession, and the specific arrangement order of the respective control rules sound. Changes in speech can largely reflect changes in mood. For example, when two people are on the phone, although they are unable to observe the facial expression and the physiological state of the other person, it is possible to roughly estimate the emotional state of the speaker by voice.

Emotion is a state that combines human feelings, thoughts, and behaviours. It includes people's psychological reactions to the outside world or their own stimuli, including the physiological reactions that accompany this psychological reaction. In people's daily work and life, the role of emotions is everywhere. In the product development process, if we can identify the emotional state of the user using the products in the process, to understand the user experience, it is possible

DOI: 10.5121/sipij.2022.13102

to improve product features, design more suited to user needs. In various human - computer interaction systems, if the system can recognize the emotional state of the person, the interaction between the person and the machine becomes more friendly and natural. Therefore, the use of computers to analyze and identify emotions, so that the machine is more humane, diverse, anthropomorphic, is of great significance. There are many useful applications for SER such as Call Centre Systems, In-car Board System, Health Care Systems, E-learning Field, and also in the Stress Management System. In this paper we implement two deep learning models LSTM and DNN for Arabic emotion recognition. The rest of this paper is structured as flow: the literature review presented in section 2 and section 3 is describes the methods and techniques. The results are detailed and discussed in section 4. Lastly, in section 5 the conclusions are presented.

2. LITERATURE REVIEW AND RELATED WORK

The speech emotion recognition referred to in this paper refers to automatic speech emotion recognition, that is, the emotion of the speaker is automatically calculated by the emotion classifier, and no human intervention is required. The author in [1] presented a new algorithm that uses the Gray Level Co-occurrence Matrix (GLCM) based Haralick feature (energy, mean, entropy and standard deviation) from spectrogram and the SVM classifier was used to recognize the emotions. For the experiments, the standard database Berlin Emotion Database and Real-time speech samples were used and for enhancing the collected real data in noisy environment adaptive filtering with Least Mean Square (LMS) algorithm was used.

An approach for the emotional state of a speaker classification was proposed by [2]. This approach based on extracting acoustic features from small speech terms and training deep recurrent neural network. The experiments were performed over the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database. The One-label approach and Connectionist Temporal Classification (CTC) approach was used to training the network. The experiments were based on comparing between three classification approaches Frame wise classification, One-label approach and CTC approach and the third approach outperforms over the other approaches. And they conclude that the emotional information may exist in few frames in the utterance and one utterance may contain many emotion states.

Speech emotion recognition method using an ensemble of random deep belief networks (RDBN) was presented in [3]. It based on extracting low-level features as input to construct lots of random sub-spaces which provided for DBN to get the higher level features which are the input to the classifier to get the emotion label. Then the majority voting used to decide the final emotion label from all emotion labels. In the experiments they used four speech databases Berlin emotional speech database in German (EMODB), Surrey Audio-Visual Expressed Emotion Database (SAVEE), Speech Emotion Database of Institute of Automation Chinese Academy of Sciences (CASIA) and FAU AIBO Emotion Corpus. But their approach achieved lower accuracy on the small and less variety training speech emotion databases. And they concluded that RDBN performs better than the DBN, SVM, and KNN, and the training database must be large for containing all type of samples.

In this paper [4], a basic emotions recognition system in the Arabic speech was presented and to find the more efficient features the system is based on extracting the Pitch, Energy, MFCCs, Formant, LPC and the spectrogram from the created emotional speech database REGIM_TES. The system recognizes five emotional states: neutral, sadness, fear, anger and happiness in the Arabic speech. Their experiments results showed that using SVM for classification with RBF kernel and combining many features produce better recognition performance. And they conclude that the Arabic speech analysis requires more investigations for the best selection of descriptors, and combining several classifiers especially the fuzzy logic may improve the precision.

A comparative study between three emotion classifications methodologies was performed by [5] to enhance the recognition of the emotional state from Arabic speech signal. They built an Arabic Emotional Speech Corpus contains Happiness, Anger, Sadness, Surprise - and Neutrality emotional states. After extract the MFCC they implemented Support Vector Machines (SVM) with Radial Basis Function (RBF) kernel, Neural Networks (NNs) and Deep learning approach using Recurrent Neural Network (RNNs). And the experimental results showed the (SVM) classification approach overcome than the (NNs) and (RNNs). But the anger and surprising states have a wrong prediction. And the lowest recognition rate was achieved in Happiness emotion.

In [6] 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.

3. Methods and Techniques

In order to recognize the emotion for the speech we follow four steps: first pre-processing the signal (speech). Secondly extract the features from the signal and thirdly build and training the LSTM model. Finally testing the model to see it capability to identify the Emotion as we can see in Figure (1) bellow:

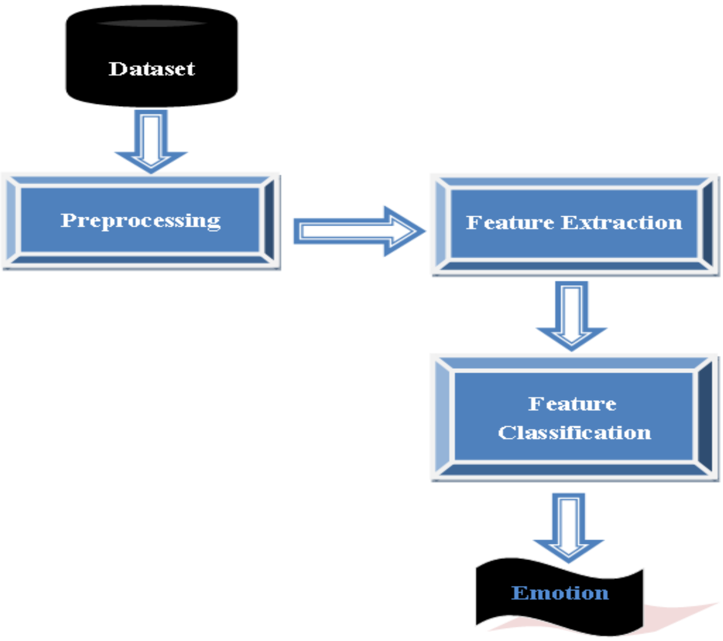

Figure 1. The Methodology of Arabic Speech Emotion Recognition System

3.1. 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.

3.2. Preprocessing

The Pre-processing stage is the first stage which aims to distinguish the voiced or unvoiced signal and create feature vectors also to adjust the speech signal so that it will be more acceptable for the next feature extraction stage. Here in this paper we apply two preprocessing algorithms: Pre-emphasis and Silent Remover.

3.3. Feature Extraction

The next step after preprocessing is the feature extraction step which it's the main part of the speech emotion recognition and its aim to extract the features from the input speech (signal) to help the system in identifying the speaker's emotion. It transforms the pre-processed speech signal to a concise but logical representation that is more discriminative and reliable than the actual signal[9]. Here in this paper five different feature types are investigated using the LSMT architecture: Mel-Frequency Cepstral Coefficients (MFCC) features, chromagram, Mel-scaled spectrogram, spectral contrast and tonal centroid features (tonnetz). The Five features are selected after reviewing several research works.

3.4. Feature Classification

After preparing the audio files in preprocessing step and extract the features from the preprocessed files in features extraction step; the role come to Feature Classification step which it aim to using the extracted features to training the model in order to classified the emotions. There are several machine learning-based classifiers that have been used by researchers to distinguish emotional classes: SVMs, RFs, and the KNN algorithm, hidden Markov models (HMMs), MLPs, and Gaussian mixture models (GMMs) [10]. In this study we use LSTM and DNN to classify. We evaluate the performance of these classifiers in terms of accuracy.

3.4.1. Long Short Term Memory (LSTM)

LSTM is the specific variant and an advanced version of a simple Recurrent Neural Networks (RNN); it learns the long-term dependencies in sequences, accordingly capturing global information over utterances. It extends the notions of a feed forward architecture by adding self connections to units as well as hidden layers at previous time steps [11]. RNNs comprise a loop, making them recurrent and permitting information to persist. Simple recurrent neural networks contain just one loop while other, more complex RNN, are composed with one or more gates allowing them to retain and forget information [12]. The success of this type of neural network is outstanding mainly to the specific variant, which are the Long Short Term Memory (LSTM). The main idea behind it is to use several gates to control the information flow from previous steps to the current steps. LSTM network was designed to model temporal sequences and their long-range dependencies more accurately than conventional RNN.

3.4.1.1. LSTM Architecture

The LSTM architecture is consists of 4 LSTM layers with cell size 100 are used, each layer is followed by dropout layers, and finally fully connected dense layers with softmax layer as the output as follow in Figure 2.

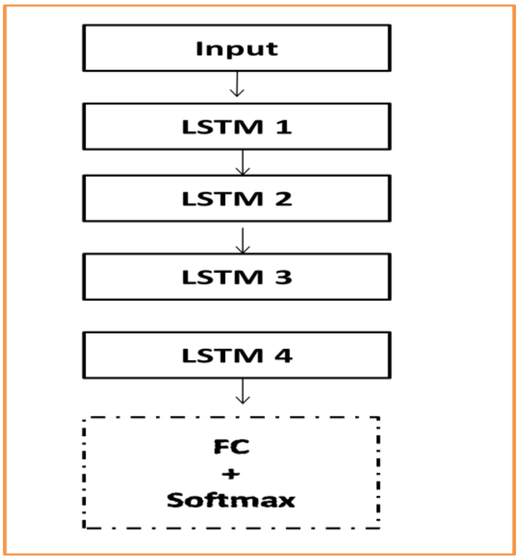

Figure 2. A stacked LSTM Layer with Fully Connected layers

4. Results and Discussion

Our proposed LSTM model was implemented and run in laptop with processor Intel(R) Core™ i7-3520 CPU @ 2.90GHz, 8GB RAM and 64 bit Operating System.

4.1. Experimental Setup

We implemented the Speech emotion classification model using Keras deep learning library with Tensorflow backend. The LSTM model were 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 16 for 200 Epoch, Categorical cross entropy loss function was used. The accuracy and loss curves for the two models are shown in Figure.3 and Figure.4.

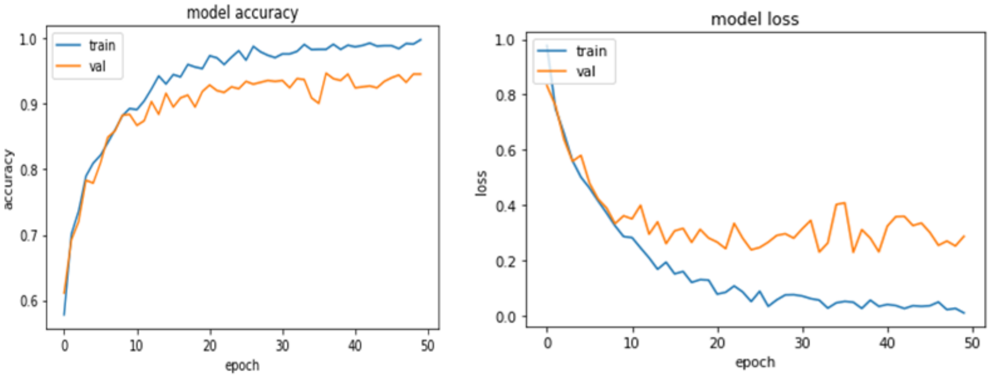

Figure 3. LSTM Model Accuracy and loss curve

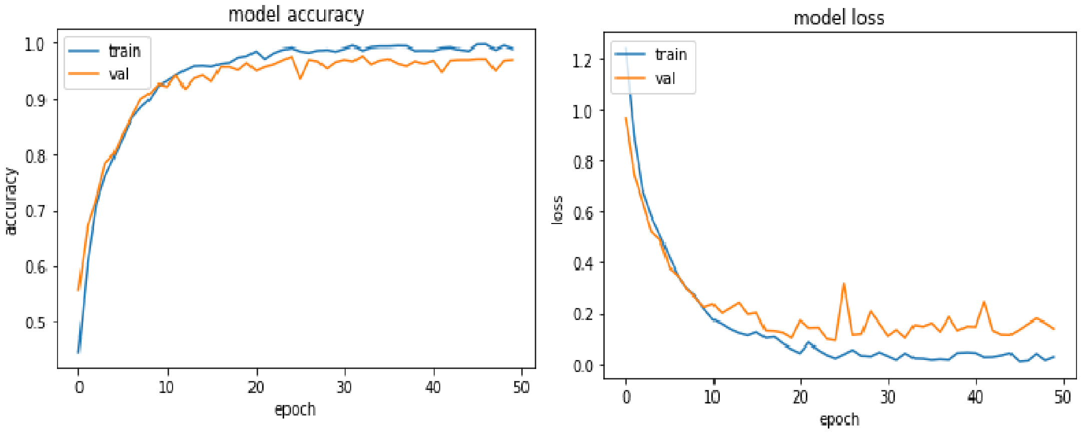

Figure 4. DNN Model Accuracy and loss curve

The details of the training the models and the achieved Accuracy is describes in Table 1 below. That result is without applying the preprocessing the accuracy is 96.81% for LSTM model and 93.34% for DNN. In order to investigation the effect of applying the pre-processing stage Table 2 shows the results after applying the two pre-processing, where they are enhanced to be 97.44% for LSTM and 97.78% for DNN . We also investigate the effect of increase number of feature to seven features as shown in Table 3, but the accuracy is decreased to 95.84% for LSTM and increase to 98.06% for DNN. Also the confusion matrix for both model are shown in Figure 5 and Figure 6.

Table 1. Training details

Classifier	**# Epoch**	**# Layers**	**# Features**	**# Emotions**	**Accuracy**
LSTM	**200**	**4**	**5** [MFCC,Chromagram, Mel-Spectrogram, Spectral Contrast And Tonnetz**]**	**4**	**96.81%**
DNN	**200**	**5**		**4**	**93.34%**

Table 2. The Accuracy after adding two preprocessing algorithms

Classifier	**# Epoch**	**# Layers**	**# Features**	**# Emotions**	**preprocessing**	**Accuracy**
LSTM	**200**	**4**	**5**	**4**	**[Pre_eemphasis and Silent Remove]**	**97.44%**
DNN	**200**	**5**	**5**	**4**		**97.78%**

Table 3. The Accuracy after adding two more Features

Classifier	**# Epoch**	**# Layers**	**# Features**	**preprocess**	**Accuracy**
LSTM	**200**	**4**	**7** [mfcc, chromagram, Mel-spectrogram, spectral contrast and tonnetz.] + **[Enrgy and Zero Crossing]**	**[Pre_eemphasis and Silent Remove]**	**95.84%**
DNN	**200**	**5**			**98.06%**

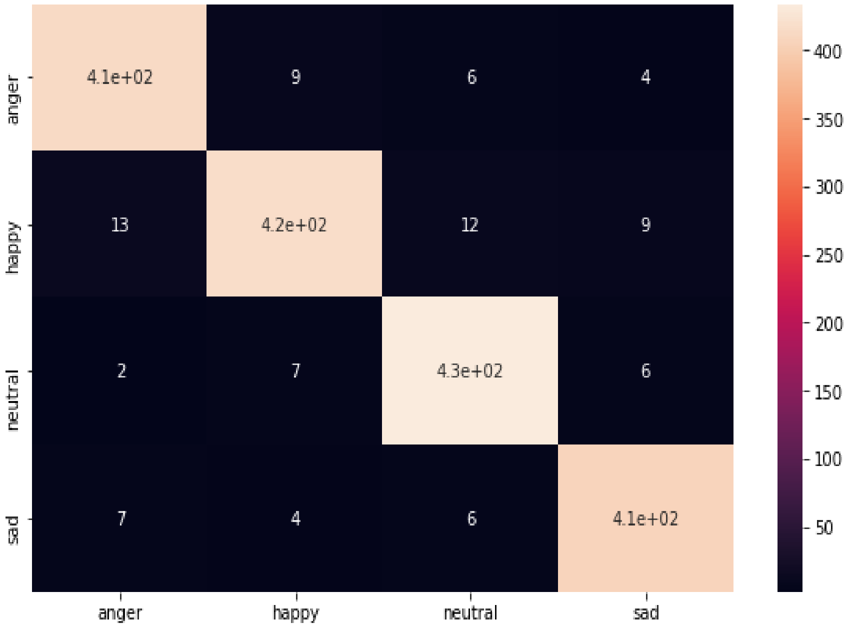

Figure 5. Confusion Matrix for LSTM Model

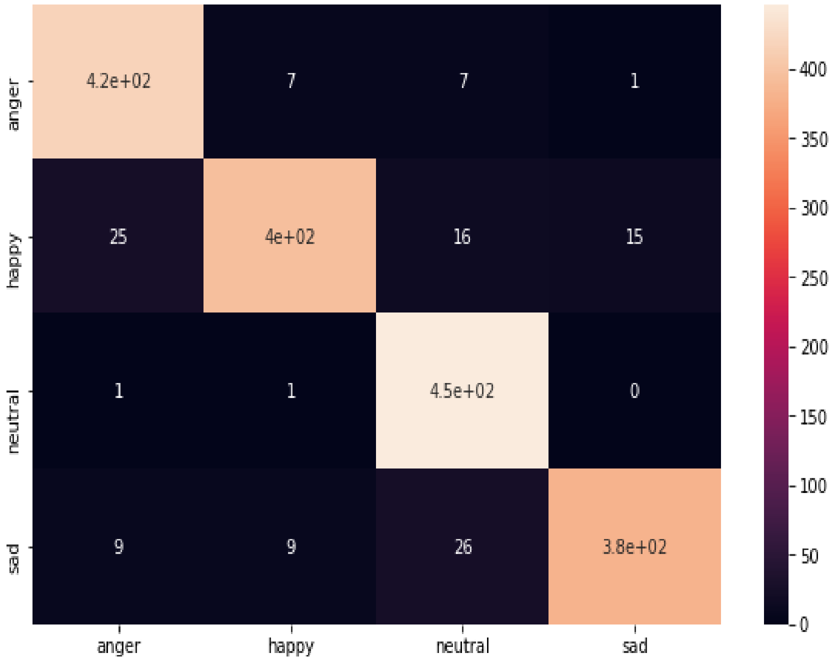

Figure 6. Confusion Matrix for DNN Model

5. Conclusion

Basic Arabic Expressive Speech corpus (BAES-DB) dataset is used in this paper to training and testing ASERS-LSTM model for Arabic Speech Emotion Recognition based on LSTM. To implement this model we using Python Programming Language and OpenCV library for programming functions with Keras and TensorFlow Open-Source Neural-Network library. We extract these five features: Mel-Frequency Cepstral Coefficients (MFCC) features, chromagram, Mel-scaled spectrogram, spectral contrast and tonal centroid features (tonnetz). We investigate the effect of applying the preprocessing by test training the model without applying the preprocessing and we get 96.81% accuracy for LSTM model and 93.34% for DNN. The results after applying the two preprocessing where there are an enhanced in the accuracy to be 97.44% for LSTM and 97.78% for DNN. We also investigate the effect of increase number of feature to seven features but the accuracy is not enhanced for LSTM it’s decreased to 95.84% and increase to 98.06% for DNN.

References

[1] L. Mathew and S. Salim, “From Speech Spectrogram Using Gray-Level Co-Occurence Matrix,” pp. 6120–6129, 2017, doi: 10.15680/IJIRCCE.2017.

[2] V. Chernykh, G. Sterling, and P. Prihodko, “Neural Networks,” 2017.

[3] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, “Random Deep Belief Networks for Recognizing Emotions from Speech Signals,” Comput. Intell. Neurosci., vol. 2017, 2017, doi: 10.1155/2017/1945630.

[4] M. Meddeb, H. Karray, and A. M. Alimi, “Automated Extraction of Features from Arabic Emotional Speech Corpus,” vol. 8, pp. 184–194, 2016.

[5] A. Al-faham and N. Ghneim, “RESEARCH ARTICLE TOWARDS ENHANCED ARABIC SPEECH EMOTION RECOGNITION : COMPARISON BETWEEN THREE METHODOLOGIES,” 2016.

[6] 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, doi: 10.1007/s10470-018-1142-4.

[7] Y. A. Ibrahim, J. C. Odiketa, and T. S. Ibiyemi, “Preprocessing technique in automatic speech recognition for human computer interaction: an overview,” Ann. Comput. Sci. Ser., vol. 15, no. 1, pp. 186–191, 2017.

[8] B. M. Nema and A. A. Abdul-Kareem, “Preprocessing signal for Speech Emotion Recognition,” Al-Mustansiriyah J. Sci., vol. 28, no. 3, p. 157, 2018, doi: 10.23851/mjs.v28i3.48.

[9] V. Picchio, V. Cammisotto, F. Pagano, R. Carnevale, and I. Chimenti, “Some Commonly Used Speech Feature Extraction Algorithms,” in Intechopen, no. Cell Interaction-Regulation of Immune Responses, Disease Development and Management Strategies, 2020, pp. 1–15.

[10] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. Bin Zikria, “Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network,” Sensors (Switzerland), vol. 20, no. 21, pp. 1–18, 2020, doi: 10.3390/s20216008.

[11] K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, “Deep Learning Techniques for Speech Emotion Recognition : A Review,” 2019 29th Int. Conf. Radioelektronika, no. July, pp. 1–6, 2019.

[12] N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, “Bidirectional deep architecture for Arabic speech recognition,” pp. 92–102, 2019.