2019 IEEE International Conference on Signal Processing, Information, Communication & Systems(SPICSCON)
28-30 November,2019, Dhaka, Bangladesh

Convolutional Neural Network (CNN) Based Speech-Emotion Recognition.

Alif Bin Abdul Qayyum, Asiful Arefeen*, Celia Shahnaz
Department of Electrical and Electronic Engineering,
Bangladesh University of Engineering and Technology, Dhaka 1205, Bangladesh.
*E-mail: asifularefeen1234@gmail.com

***Abstract*—Speech is considered as the widest and most natural medium of communication. Speech can convey a plethora of information regarding one's mental, behavioral, emotional traits. Besides, speech-emotion recognition related work can aid in averting cyber crimes. Research on speech-emotion recognition exploiting concurrent machine learning techniques has been on the peak for some time. Numerous techniques like Recurrent Neural Network (RNN), Deep Neural Network (DNN), spectral feature extraction and many more have been applied on different datasets. This paper presents a unique Convolutional Neural Network (CNN) based speech-emotion recognition system. A model is developed and fed with raw speech from specific dataset for training, classification and testing purposes with the help of high end GPU. Finally, it comes out with a convincing accuracy of 83.61% which is better compared to any other similar task on this dataset by a large margin. This work will be influential in developing conversational and social robots and allocating all the nuances of their sentiments.**

***Index Terms*—Recurrent Neural Network (RNN), Deep Neural Network (DNN), Convolutional Neural Network (CNN), Speech Processing, Emotion, Speech-emotion, SAVEE dataset, Emotional state, Discrete Cosine Transform (DCT), Modulation Spectral Features (MSF), Support Vector Machines (SVM), Multivariate Linear Regression Classification (MLR), Multi class.**

I. Introduction

'Emotion' is a strong feeling deriving from one's circumstances, mood, or relationships with others and an integral part in human life. It can be of many types and interchanges in seconds. It is closely related to our decision making approach. Since human-computer interaction depends much on user's emotional state and other nuances, it can be made more flexible and interactional if our state of mind can be accurately determined. Furthermore, emotion recognition can be useful in inquisition process. Low requirements of heavy duty hardware setup and complex computational algorithm make research in this field very admirable among the researchers.

With the recent boom in machine learning backed speech processing, numerous tools are being used in emotion recognition. Some researches are performed under real-life scenarios to make it more realistic and convince their practical viability. As a result, one of the most important sides of these researches i.e. the accuracy is changing vastly for adopting different methods.

Many research works have been performed in order to determine instantaneous human emotions using his or her speech. Several machine learning based approaches have been responsible for the discrepancy in their obtained results and performances. Rieger et. al. [1] have emphasized speech-emotion recognition based on spectral feature extraction and an ensemble of k nearest neighbor (kNN) classifiers.

Kerkeni et. al. [2] have employed recurrent neural network (RNN) classifier to classify seven emotions first and achieved an accuracy of 83% on Berlin dataset when a speaker normalization (SN) and a feature selection are applied to the features.

Tripathi et. al. [3] from Samsung R&D Institute India exploited combined MFCC-Text Convolutional Neural Network (CNN) model which proved to be highly accurate in recognizing emotions in IEMOCAP data.

Tao et. al. [4] propose a practically feasible work by providing ensemble framework which can capture several aspects of characteristics related to emotion. The framework is evaluated on multimodal emotion challenge (MEC) 2017 corpus. The corpus is collected from Chinese films and TV programs, whose scenarios are close to the real world.

Sidorov et. al. [5] tried to figure out the most essential features with self-adaptive multi-objective genetic algorithm as a feature selection technique and a probabilistic neural network as a classifier. The proposed approach was evaluated using a number of multi-language databases (English, German), which were represented by 37- and 384-dimensional feature sets.

Sethu et. al. [6] examine current approaches to speech based emotion recognition and provide a comparative image on the numerous methods ongoing.

In this paper, a CNN based emotion classification has been presented that does not require any preprocessing of the input data stream and that is computationally efficient. The singularity of this work is that it has achieved superiority in terms of accuracy over other concurrent and related works.

In the following paragraphs, the used database, followed methodology and obtained results are illuminated. A comparative analysis of the accuracy is also presented finally.

II. Database

SAVEE(Surrey Audio-Visual Expressed Emotion) [7], [8] database has been recorded from four native English male speakers, postgraduate students and researchers at the University of Surrey aged from 27 to 31 years. Emotion has been described psychologically in discrete categories: *Anger, Disgust, Fear, Happiness, Sadness* and *Surprise*. This is supported by the cross-cultural studies of Ekman [9] and studies of automatic emotion recognition tended to focus on recognizing

978-1-7281-4294-4/19/$31.00 ©2019 IEEE

these. In the Kaggle version of this dataset, *Neutral* class has been added to provide recordings of 7 emotion categories. The text material consists of 15 TIMIT sentences per emotion: 3 common, 2 emotion-specific and 10 generic sentences that are different for each emotion and phonetically-balanced. The 3 common and $2 \times 6 = 12$ emotion-specific sentences have been recorded as Neutral to give 30 Neutral sentences. This results in a total of 120 utterances per speaker.

The data were recorded in a visual media lab with high quality audio-visual equipment, processed and labeled. To check the quality of performance, the recordings were evaluated by 10 subjects under audio, visual and audio-visual conditions.

III. Methodology

The followed methodology can be discussed in the following points-

A. Data Preprocessing

Two completely different approaches have been compared for this purpose. In the feature extraction based process, modulation spectral features and MFCC have been selected to extract the emotional features just as mentioned in [2]. For each frame, the Fourier transform and the energy spectrum have been estimated and mapped into the Mel-frequency scale. The discrete cosine transform (DCT) of the Mel log energies is estimated and the first 12 DCT coefficients provides the MFCC values used in the classification process. First 12 order of the MFCC coefficients have been extracted where the speech signals' sampling frequency is 44.1 KHz. For each order coefficient, the mean, variance, standard deviation, kurtosis, and skewness have been calculated and they are the same for all other frames of an utterance. Each MFCC feature vector is 60-dimensional. Modulation Spectral Features (MSF) are obtained by emulating the spectro-temporal (ST) processing performed in the human auditory system and consider regular acoustic frequency jointly with modulation frequency. Speech signal is first decomposed by an auditory filter bank (19 filters in total). The Hilbert envelopes of the critical-band outputs are computed to form the modulation signals. A modulation filter bank is further applied to the Hilbert envelopes. The spectral contents of the modulation signals are referred to as modulation spectra, and the proposed features are thereby named Modulation Spectral Features (MSFs) [11]. In this experiment, an auditory filterbank with N=19 filters and a modulation filter bank with M=5 filters are used. In total, 95 (19 x 5) MSFs are calculated in this work from the ST representation. Both the MFCC and ST computation methods are done in the same process mentioned in [2]. The effect of speaker normalization (SN) is applied, which removes the mean of features and normalizes them to unit variance. Experiments are performed under a speaker-independent condition.

MFCC and MSF computation methods are shown in Fig. 1 and Fig. 2 respectively.

For this proposed Deep Convolutional Neural Network (CNN) method, there is no such need to preprocess the data. To train the model the raw audio data is used by directly feeding the neural network model. For training purpose the data is split into three different parts- Train, Validation and Test dataset- as shown on Table I.

TABLE I: Data allocation.

Dataset Type	Train	Validation	Test
Audio Sample Number	300	100	80

B. Recursive Feature Elimination (RFE)

A model (e.g., linear regression or SVM) has been used to select either the best- or worst- performing feature and then excludes this feature. These estimators assign weights to features (e.g., the coefficients of a linear Model). First, the estimator is trained on the initial set of features, and the predictive power of each feature is measured [12]. Then, the least important features are removed from the current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached. In this work, the recursive feature elimination method of feature ranking have been implemented via the use of basic linear regression (LR-RFE) [13]. Like the feature computation, the feature selection method has also been followed just as mentioned in [2].

C. Feature Based Classification Methods

Three different classifiers have been chosen: Multivariate Linear Regression Classification (MLR) [14], Support Vector Machines (SVM) [10,13,14,40,41] and Recurrent Neural Networks (RNN) [15].

D. Model Architecture of Deep Convolutional Neural Network

This model has total 7 one-dimensional convolutional layers, each followed by a batch normalization layer and a max pooling layer with a pool size of 2, except for the last convolutional layer. Input to the model was raw audio with duration of 8 seconds. The audio files with duration less than 8 seconds was zero padded. Number of filters in convolutional layers are 32, 64, 128, 256, 512, 1024 and 1024 accordingly. Kernel sizes of those filters are 21, 19, 17, 15, 13, 11 and 9 accordingly. The last one-dimensional convolutional layer is followed by a global max pooling layer and then followed by a dense layer with 128 nodes. Activation functions for all the convolutional and dense layers so far is 'Relu' [16]. The last layer of the model is a dense layer with 7 nodes (As the total number of emotional classes for the dataset is 7) with activation function 'softmax' [17]. Model architecture is shown in Fig. 3.

E. Training CNN model

In this paper, a deep convolutional neural network has been suggested to classify emotions from speech signal from the SAVEE dataset. To train the model, Adam [18] optimizer was used with an initial learning rate of 0.001, beta_1 value of 0.9 and beta_2 value of 0.999. Loss function used for this purpose is 'Categorical Crossentropy'. The model was trained over 400 epochs. Training and validation loss and accuracy curve is shown in Fig. 4 and Fig. 5 respectively.

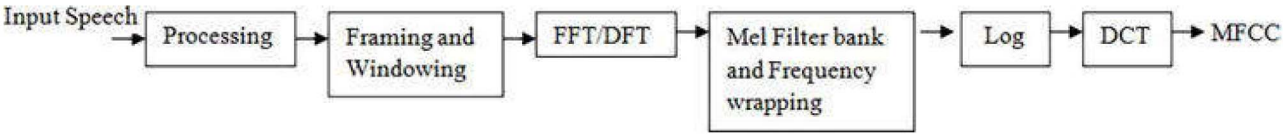

Fig. 1: Schema of MFCC extraction [2].

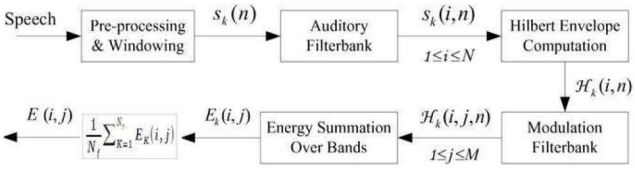

Fig. 2: Process for computing the ST representation [2].

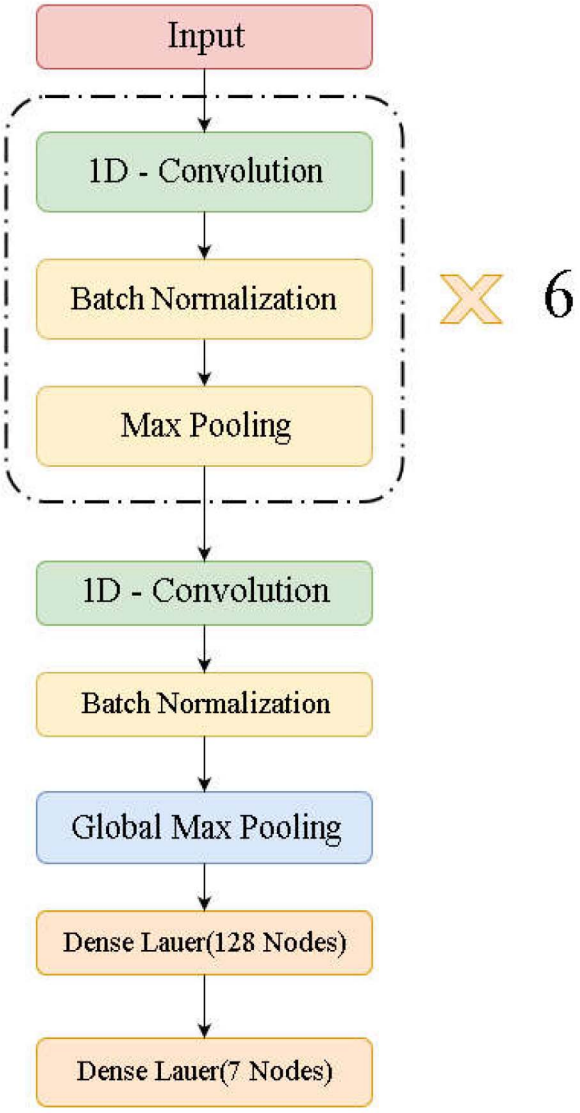

Fig. 3: Schematic diagram of the model.

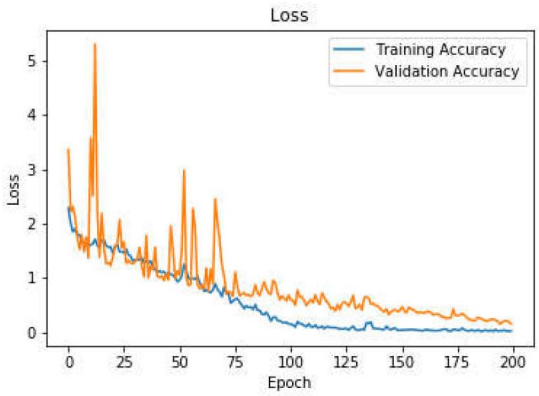

Fig. 4: Graph for validation loss.

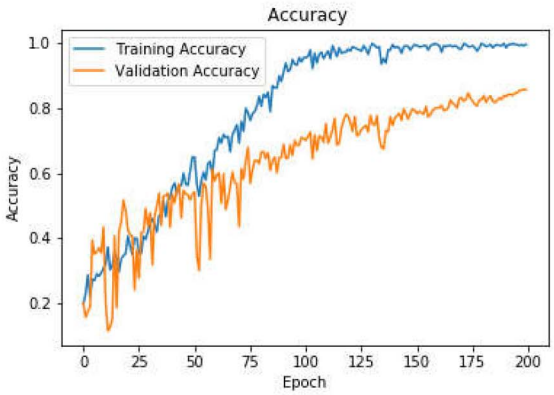

Fig. 5: Graph for validation accuracy.

IV. Results

A. Comparison Between Different methods in Accuracy

Table II summarizes the results for different methods. Here, four methods (MVR, SVM, RNN and CNN have been applied on the test data which constitutes a total of 80 audio files, 20 audio files from each speaker. In that table, the classes : 'A', 'D', 'F', 'H', 'N', 'SA' and 'SU' stand for anger, disgust, fear, happiness, neutral, sadness and surprise respectively. From the table it can be seen that Deep Convolutional neural network outperforms all the other three methods.

Table III summarizes the performance of the CNN model on the test dataset. From the table it can be concluded that the model performs almost uniformly on all the classes of emotions. Fig. 6 also shows the normalized confusion matrix for the CNN model applied on the test dataset.

TABLE II: Accuracies for MVR, SVM, RNN and proposed CNN methods.

Feature	Method	A	D	F	H	N	SA	SU	AVG
MFCC	MVR	45.68	45.44	42.79	67.08	59.43	69.91	71.94	57.47
MS		41.93	41.03	39.97	68.97	52.17	61.11	69.87	53.58
MFCC+MS		58.47	57.59	60.04	75.13	68.74	79.26	83.07	68.90
MFCC	SVM	53.17	57.31	48.97	79.81	68.34	78.19	81.48	66.75
MS		49.95	52.31	49.79	77.79	63.71	72.13	78.91	63.51
MFCC+MS		69.87	72.69	71.41	78.18	75.47	82.63	85.14	76.48
MFCC	RNN	61.87	57.47	53.98	74.81	68.31	78.87	82.67	68.28
MS		53.85	61.13	52.36	82.69	78.22	72.35	78.84	68.49
MFCC+MS		70.32	75.22	77.54	78.19	79.47	87.17	86.47	79.20
	CNN	73.50	80.88	80.12	87.38	86.38	89.75	87.25	83.61

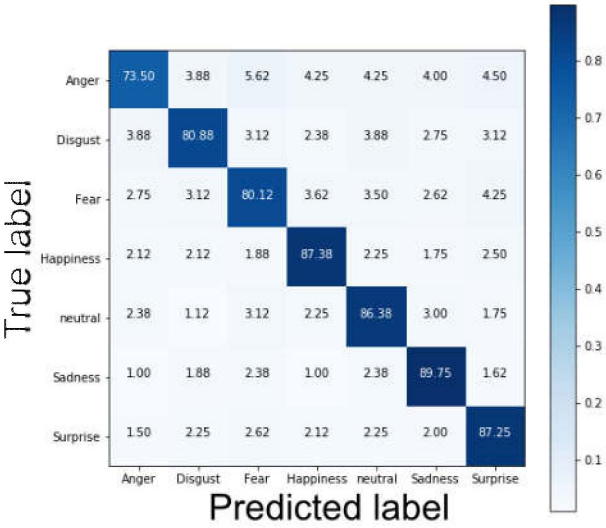

Fig. 6: Confusion matrix for CNN method.

TABLE III: Recognition results for the Deep convolutional Neural Network method on Test data.

Class	Precision	Recall	F1-score
Anger	0.85	0.73	0.79
Disgust	0.82	0.81	0.82
Fear	0.85	0.80	0.82
Happiness	0.83	0.87	0.85
Neutral	0.83	0.86	0.85
Sadness	0.83	0.90	0.86
Surprise	0.84	0.87	0.86
Total	0.84	0.84	0.84

V. Conclusion

This paper presents a novel, singular, neural entwork based speech-emotion recognition procedure that is a credible alternative to the other traditional methods. It can classify 7 types of emotion with great accuracy. It can be interpreted as a speech processing tool and can be traind to classify more types. Since it does not require any complex computational idea, it can be considered as user friendly and easily understandable. This work will have far reaching consequences in the speech-emotion recognition field and can be considered as a viable tool for future works on several pertinent sectors. It is indeed a modern way to apply speech processing and neural network in practical for the betterment of technology.

References

[1] S. A. Rieger and R. Muraleedharan and R. P. Ramachandran, "Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers," *The 9th International Symposium on Chinese Spoken Language Processing*, pp. 589-593, Sep, 2014.

[2] Kerkeni, Leila and Serrestou, Youssef and Raoof, Kosai and Cléder, Catherine and Mahjoub, Mohamed and Mbarki, Mohamed, "Automatic Speech Emotion Recognition Using Machine Learning," March, 2019.

[3] Tripathi, Suraj and Kumar, Abhay and Ramesh, Abhiram and Singh, Chirag and Yenigalla, Promod," *IEEE Trans. On Industry Applications*, vol. IA-20, no. 4, pp. 727-734, July 1984.

[4] F. Tao and G. Liu and Q. Zhao, "An Ensemble Framework of Voice-Based Emotion Recognition System," *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pp. 1-6, May 2018.

[5] Sidorov, Maxim and Brester, Christina and Minker, Wolfgang and Semenkin, Eugene, "Speech-Based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm," *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3481-3485, May 2014.

[6] Sethu, Vidhyasaharan and Epps, Julien and Ambikairajah, Eliathamby, "Speech Based Emotion Recognition," pp. 197-228, September 2015.

[7] QWang, Wenwu, "Machine Audition: Principles, Algorithms and Systems," *IGI Global, 2011*, pp. 1-554, 2011.

[8] Jackson, Philip and ul haq, Sana, "Surrey Audio-Visual Expressed Emotion (SAVEE) database," April 2011. [Online]. Available at: Surrey Audio-Visual Expressed Emotion (SAVEE) Database.

[9] Ekman, Paul, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, pp. 45-60, 1999.

[10] Melki, Gabriella and Kecman, Vojislav and Ventura, Sebastian and Cano, Alberto, "OLLAWV: online learning algorithm using worst-violators," *Applied Soft Computing*, vol. 66, pp. 384-393, 2018.

[11] Siqing Wu and Tiago H. Falk and Wai-Yip Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, pp. 768-785, 2011.

[12] Duan, Kai-Bo and Rajapakse, Jagath C and Wang, Haiying and Azuaje, Francisco, "Multiple SVM-RFE for gene selection in cancer classification with expression data," *IEEE transactions on nanobioscience*, vol. 4, no. 3, pp. 228-234, 2005.

[13] Pedregosa, Fabian and Varoquaux, Gael and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent and others, "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, vol. 12, pp. 2825-2830, October 2011.

[14] Imran Naseem and Roberto Togneri and Mohammed Bennamoun, "Linear Regression for Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 2106-2112, 2010.

[15] Srinivas Parthasarathy and Ivan Tashev, "Convolutional Neural Network Techniques for Speech Emotion Recognition," *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 121-125, 2018.

[16] Agarap, Abien Fred, "Deep learning using rectified linear units (relu)," March 2015.

[17] Nwankpa, Chigozie and Ijomah, Winifred and Gachagan, Anthony and Marshall, Stephen, "Activation functions: Comparison of trends in practice and research for deep learning," November 2018.

[18] Kingma, Diederik and Ba, Jimmy, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, December 2009.