A Hybrid Technique using CNN+LSTM for Speech Emotion Recognition

Hafsa Qazi, Baij Nath Kaushik

Abstract: ***Automatic speech emotion recognition is a very necessary activity for effective human-computer interaction. This paper is motivated by using spectrograms as inputs to the hybrid deep convolutional LSTM for speech emotion recognition. In this study, we trained our proposed model using four convolutional layers for high-level feature extraction from input spectrograms, LSTM layer for accumulating long-term dependencies and finally two dense layers. Experimental results on the SAVEE database shows promising performance. Our proposed model is highly capable as it obtained an accuracy of 94.26%.***

Keywords: ***CNN, LSTM, RNN, SER, Spectrograms.***

I. INTRODUCTION

We may ask ourselves why the emotional awareness by machines is even desirable [1]. Firstly to make the customer experience better by understanding their emotional state for example in a lot of computer-aided learning systems, we can adjust the presentation of the material or pace of learning by knowing the learner's emotional state and achieve the best results for that student [2]. Emotional awareness by machines can also be used to provide tools to humans to make them more effective for example in gaming industry we can look or hear the gamer's reaction and improve the game design by knowing the frustration point that is there in the game design. In commercial marketing, we can use emotion recognition to gauge the viewers' reaction to the marketing materials and accordingly fine-tune the materials to achieve the desired effect [3], [4]. Therefore, in all these applications, we need to build machines that are capable to perceive human emotional state. People perceive emotion from speech, facial expressions, body language etc., and speech being the most natural and fastest way [5]. Emotion perception from speech is not that simple, this is because in most of the cases we actually rely on the context (in longer conversations to get better sense). For many years, SER has been a dynamic research area [6], [7]. Emotion representation theory became the foundation for emotion recognition research by providing methods to give the various details of emotions so as to label the data with appropriate target and finally the machines can learn to predict the emotions [8].

Earlier, a lot of scientists focussed on finding the best feature that represents emotion. Our understanding of what signal level information in speech is most useful (for emotion recognition) improved only by the advances that scientists did in classical machine learning and signal processing [9] [10]. In recent years, with the popularity of deep learning in fields such as computer vision or speech recognition, emotion recognition also pivoted into this new deep learning technology based approach [11]. The remainder of this paper is arranged as follows: a brief review of current related work is presented in the next section, following the details of our proposed CNN-LSTM based speech emotion recognition system. The experimental results are also discussed and finally the conclusion.

II. RELATED WORK

This section gives an overview of the recent trends in speech emotion recognition systems. At present, two mostly used deep learning methods are Convolution Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. Recently, researchers used both CNN and LSTM based systems to advance the performance of SER systems using spectrogram-based speech signals or raw speech signals [12], [13], [14]. Huang et al. [15] used auto-encoders followed by single-layer Convolution Neural Network (CNN) for speech emotion recognition and achieved good performance. However this approach is not suitable for variable length speech inputs. Sainath et al. [16] used CLDNN (Convolution Long Short-Term Memory Deep Neural Networks) on raw waveform speech signals. Tregeorgis et al. [17] integrated two-layer CNNs with Long Short Term Memory (LSTM) and presented an end-to-end speech emotion recognition system. X. Li et al. [18] presented a large vocabulary speech recognition using CNN and LSTM-RNN. Badshah et al. [19] proposed a three-layer CNNs using spectrograms of speech signals and reconstructed error-based framework for continuous data. Mirsamadi et al. [20] and Han et al. [21] proposed Recurrent Neural Networks (RNNs) along with LSTM for speech emotion recognition system. Nie et al. [22] proposed deep retinal CNNs which achieved 99.25% accuracy. Shiqing et al. [23] proposed CNN-LSTM based model using two challenging spontaneous databases – AFEW5.0 and BAUM-1s that outperformed state-of-the-art methods.

Revised Manuscript Received on June 15, 2020.
* Correspondence Author

Hafsa Qazi*, Department of Computer Science, Shri Mata Vaishno Devi University, Katra, J&K, India.

Baij Nath Kaushik, Associate Professor, Department of Computer Science & Engineering, SMVDU, Katra, J&K, India.

III. PROPOSED METHODOLOGY

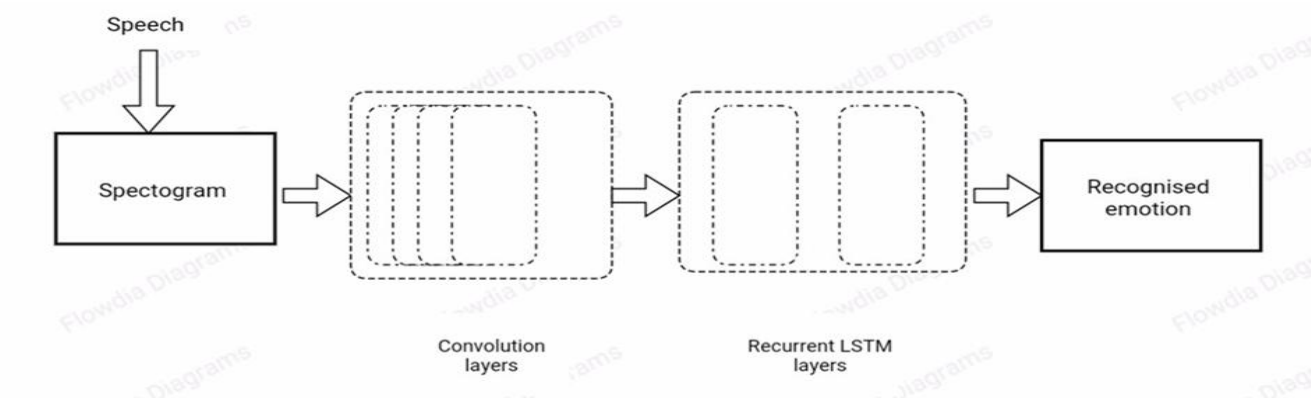


Figure 1: Architecture of our proposed model

Figure 1 depicts the basic architecture of our work. It broadly includes two parts: (1) creation of spectrograms of the speech signal, (2) network architecture: CNN-LSTM fusion. This fusion is very effective as it takes the advantage of both these neural networks.

A. Creation of spectrograms

The visual representation of sound is called a spectrogram, fashioned by a mathematical algorithm called fast-fourier transform. A raw speech signal is taken and decomposed it into its frequency components by using this algorithm [24]. Simply put, a spectrogram reflects the variation of frequency in the signal [25]. A spectrogram displays horizontal x-axis time and vertical y-axis frequency. The components (which form a complex signal) in speech signal do not have the same amplitude value. Differences in the amplitude are shown on a spectrogram by shading [26]. In this way, a spectrogram is three dimensional i.e. it shows x-axis time, y-axis frequency and shading amplitude similar to RGB [27]. Sample spectrograms are shown in figure 2. In our model, we have used 480 utterances whose corresponding spectrograms are generated by using python's spectrogram function from pyplot library.

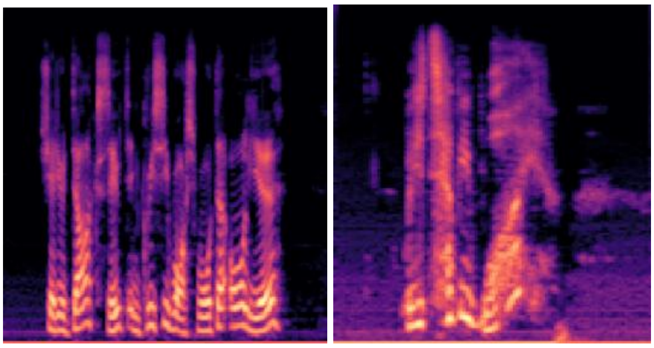


Figure 2: Spectrograms generated from speech inputs.

B. Network Architecture: CNN-LSTM Fusion

We will briefly discuss the algorithms and the layers used in our model and then throw light on CNN-LSTM fusion.

1) Convolution neural network

In 1998, Yann LeCun introduced the Convolution Neural Network (CNN). It is currently used in so many applications from the classification of images to audio synthesis [28]. Most commonly, we should think of a convolution neural network as an artificial neural network that has some ability to identify patterns and make sense of them. This pattern detection makes neural networks such useful for image analysis [29]. There are mainly four layers in CNN and we use some additional layers in order to normalize our network. Each of these are described below:

- **Convolution layer:** The convolution layer represents CNN's layer one where we interact (images, 1D time series data) with filters or kernels. By using a sliding window, we apply small units across the input and these units are known as filters [30]. The input and filter depth are synonymous, a coloured image RGB having depth three, will be filtered with same depth i.e. three [31]. In convolution process, element-wise product of filters in the image is taken and then for each sliding action the products are added. We will obtain a 2-D matrix after convolving a 3-D filter as the output [32].
- **Activation layer:** Between the successive convolutional layers, we only use non-linear activation functions [33]. Due to associative property of convolution, only nonlinear activation functions between successive convolution layers are allowed and linear activation functions do not lead to learning.
- **Pooling layer:** Pooling includes the down-sampling of the features with the goal that we have to learn less parameters during training [34]. Using pooling layer, mainly two hyper-parameters are introduced, one being the spatial extent dimension and the other being stride. The value of 'n' defines the dimension of spatial extent, taking n * n feature representation and mapping to a single value. The number of feature the sliding window skips along the width and height is the stride. Over-fitting is reduced by performing pooling as it reduces the number of parameters [35], [36]. A 2 * 2 max non-overlapping filter having a stride of 2 represents a common pooling layer. If a maximum value is returned among the features in the region, it represents a max filter, however if the return is the average of features then it is average filter. In practice, a max filter performs better.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

- **Fully connected layer:** The high-level features in data are represented by the output of the convolution layer. We use this layer for classification.
  A fully connected layer is introduced to allow output to be flattened and connected to the output layer, to learn these features in non-linear combinations. The pooling layers' output is 3D volume but a fully connected feed forward network takes a 1D feature vector as input [37]. To convert this 3D volume into one dimension, the output width and height should be one and this is possible only by flattening the 3D layer into 1D vector [38].
- **Batch Normalization layer (Batch Norm):** During training, if there is any instability in any layer of our neural network, we apply batch normalization to that layer. The output from the activation function is normalized using a batch normalization layer and this being the very first thing that this layer does. This addition can greatly increase the training speed [39]. Also the outlying large weights greatly influence the training process and batch norm reduces it. The Batch norm name is given because it works according to per batch basis and the batch size is set when we train our model [40].

2) Long Short-Term Memory (LSTM*)*

RNN (Recurrent Neural Network) is a neural network in which the previous steps' output are fed as input to the current step. But practically, these recurrent neural networks have a limitation that they can look back only a few steps. For understanding problems such as speech recognition, a system is required to store and use context information [41]. A type of RNN is LSTM. A neural network that learns order dependencies in sequence prediction problems are known as LSTM networks. In order to rectify the vanishing gradient problem, endless efforts were employed and LSTM is one such solution [42]. The speech signal is continuous in the time domain, so that each frame function only represents the emotional characteristics in a single frame. LSTM increases the information between adjacent frames which helps to reflect temporal continuity of the features. Therefore, LSTM obviously supports speech recognition.

3) Fusion of CNN-LSTM

A convolutional neural network being a feed-forward network, filters spatial data whereas the recurrent neural network (LSTM) feeds data back into itself. Thus recurrent neural networks are better suited for sequential data [43]. Put it differently, a convolutional neural network is able to perceive patterns across space, LSTM can see them over time. Since our speech signal is sequential, so LSTM is best suited for speech processing.

C. Experimental Setup

In our model, we implemented a hybrid CNN-LSTM network to classify emotions from speech signals. We used SAVEE dataset for our study. The steps which we performed are summarised as follows: firstly, we take a speech dataset in which speech signal is classified into number of attributes like happiness, anger, sadness which we want to identify. The speech signal is then ready for processing in the frequency domain and thus the spectrogram is obtained. These spectrograms can be treated as an image. CNN does prediction when this spectrogram is fed to it.Further LSTM is used to improvise the results as they are best suited for sequential speech data. Then we test and train our architecture and finally we get the outcome i.e. recognised emotions.

- Emotional Database: The SAVEE (Surrey Audio-Visual Expressed Emotion), a public British English database having seven emotion viz. happiness, anger, disgust, fear, surprise, sadness and neutral [44]. In this database 15 utterances for each emotion are produced by four male actors. These utterances are spoken in English language. Common to all emotions were three of these utterance while emotion-specific were two. The generic sentences that were dissimilar across six emotions were represented by the remaining utterances. Neutral emotions reported using three typical and two each emotion-specific sentences. 44 kHz was the sampling rate of the recordings. Except neutral which has 120 utterances, each emotion has 60 utterances from a total of 480 utterances. In our model we have used all of these 480 utterances and we have labelled each of these according to the different emotions. Label 0-59 for anger, 60-119 for disgust, 120-179 for fear, 180-239 for happiness, 240-359 for neutral and 360-479 for sadness.

D. Training and testing sets

Our model was trained using 9-fold cross validation i.e. 9-fold data partition, first fold representing test set while others were used to train our model. Then second fold was used for testing and all other remaining to train our models. This is repeated for the third fold being the test set and process goes on for all the folds. In other words, 10 percent of the dataset is employed for testing and 90 percent of the dataset for training. Because of the computation and time expenses. The number of training epochs ware set to 100.

E. Model Architecture

Deep neural network with four convolution layers, LSTM layer and finally two dense layers is taken as the model architecture for our study. Batch normalization layer, activation layer and finally max pooling layer are next to each convolution layer. In the convolution layer, rectified linear units (ReLU) activation functions were used and in the activation layer, Exponential Linear Unit (ELU) activation was used. A 3 * 3 kernel size is employed for the convolution layer and a 2 * 2 kernel size of max pooling layer. The convolution layers' filter sizes varies as follows- filter size of 32 for 1st convolution layer, filter size of 64 for 2nd convolution layer, filter size of 128 for 3rd and 4th convolution layers. Then we resize our output before feeding it into LSTM layer as CNN works on 4D while LSTM works on 2D. And finally we used the dense layer which is used for classification.

IV. RESULTS AND ANALYSIS

We commenced our study by implementation of CNN-LSTM architecture. Subsequently, hyper-parameters were modified by changing the convolution kernel size etc.

The model is constructed in Python and trained for 100 epochs. Accuracy, recall, precision, F1 score and Cohens kappa are the parameters on which we have shown the performance analysis of our model.
The graph below depicts the results based on these parameters. Succeeding the performance graph are the screenshots of the spectrograms generated for each audio signal, the program results showing output and the emotion recognised for a particular audio input.

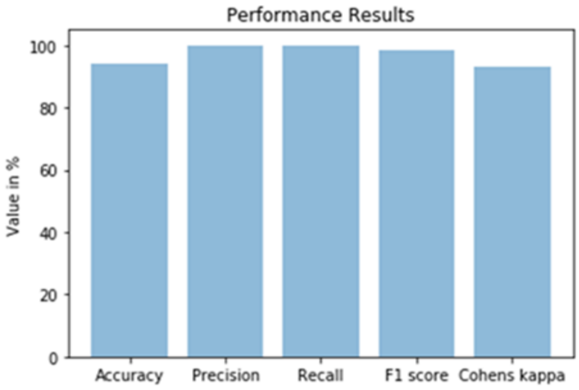
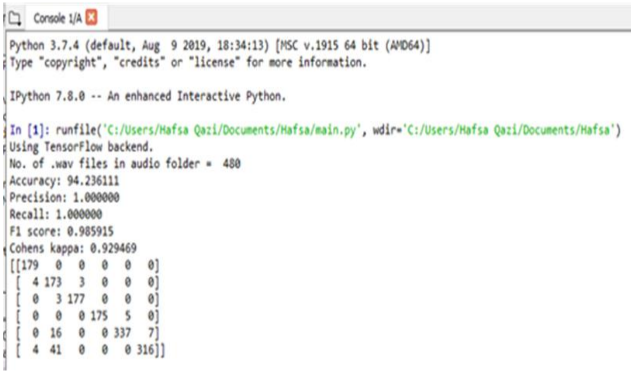


Figure 3: Performance results of our proposed model based on different parameters



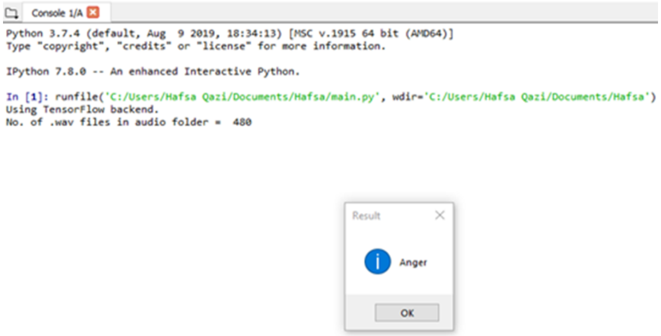
```
Console 1/A
Python 3.7.4 (default, Aug  9 2019, 18:34:13) [MSC v.1915 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 7.8.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/Hafsa Qazi/Documents/Hafsa/main.py', wdir='C:/Users/Hafsa Qazi/Documents/Hafsa')
Using TensorFlow backend.
No. of .wav files in audio folder =  480
Accuracy: 94.236111
Precision: 1.000000
Recall: 1.000000
F1 score: 0.985915
Cohens kappa: 0.929469
[[179   0   0   0   0   0]
 [  4 173   3   0   0   0]
 [  0   3 177   0   0   0]
 [  0   0   0 175   5   0]
 [  0  16   0   0 337   7]
 [  4  41   0   0   0 316]]
```

Figure 4: Screenshot of program output showing performance parameters.

Figure 4 shows the results of various parameters. Our model shows accuracy of 94.26%, precision value 1.0, recall value 1.0, F1 score 0.985.

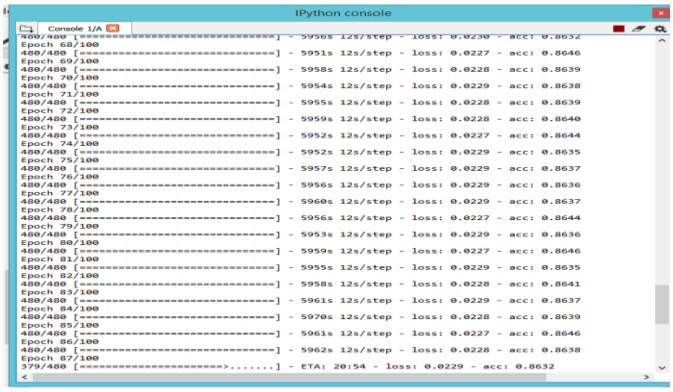


```
Console 1/A
Python 3.7.4 (default, Aug  9 2019, 18:34:13) [MSC v.1915 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 7.8.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/Hafsa Qazi/Documents/Hafsa/main.py', wdir='C:/Users/Hafsa Qazi/Documents/Hafsa')
Using TensorFlow backend.
No. of .wav files in audio folder =  480
```

Figure 5: Screenshot of program output showing the anger emotion recognised for audio input.



```
480/480 [==============================] - 5956s 12s/step - loss: 0.0230 - acc: 0.8632
Epoch 68/100
480/480 [==============================] - 5951s 12s/step - loss: 0.0227 - acc: 0.8646
Epoch 69/100
480/480 [==============================] - 5958s 12s/step - loss: 0.0228 - acc: 0.8639
Epoch 70/100
480/480 [==============================] - 5954s 12s/step - loss: 0.0229 - acc: 0.8638
Epoch 71/100
480/480 [==============================] - 5955s 12s/step - loss: 0.0228 - acc: 0.8639
Epoch 72/100
480/480 [==============================] - 5959s 12s/step - loss: 0.0228 - acc: 0.8640
Epoch 73/100
480/480 [==============================] - 5952s 12s/step - loss: 0.0227 - acc: 0.8644
Epoch 74/100
480/480 [==============================] - 5952s 12s/step - loss: 0.0229 - acc: 0.8635
Epoch 75/100
480/480 [==============================] - 5957s 12s/step - loss: 0.0229 - acc: 0.8637
Epoch 76/100
480/480 [==============================] - 5956s 12s/step - loss: 0.0229 - acc: 0.8636
Epoch 77/100
480/480 [==============================] - 5960s 12s/step - loss: 0.0229 - acc: 0.8637
Epoch 78/100
480/480 [==============================] - 5956s 12s/step - loss: 0.0227 - acc: 0.8644
Epoch 79/100
480/480 [==============================] - 5953s 12s/step - loss: 0.0229 - acc: 0.8636
Epoch 80/100
480/480 [==============================] - 5959s 12s/step - loss: 0.0227 - acc: 0.8646
Epoch 81/100
480/480 [==============================] - 5955s 12s/step - loss: 0.0229 - acc: 0.8635
Epoch 82/100
480/480 [==============================] - 5958s 12s/step - loss: 0.0228 - acc: 0.8641
Epoch 83/100
480/480 [==============================] - 5961s 12s/step - loss: 0.0229 - acc: 0.8637
Epoch 84/100
480/480 [==============================] - 5970s 12s/step - loss: 0.0228 - acc: 0.8639
Epoch 85/100
480/480 [==============================] - 5961s 12s/step - loss: 0.0227 - acc: 0.8646
Epoch 86/100
480/480 [==============================] - 5962s 12s/step - loss: 0.0228 - acc: 0.8638
Epoch 87/100
379/480 [======================>.......] - ETA: 20:54 - loss: 0.0229 - acc: 0.8632
```

Figure 6: Screenshot of epochs running python console

V. CONCLUSION

In this study, we applied CNN-LSTM fusion to characterize emotional states of acted speech utterances using SAVEE dataset. The objective here is to propose a CNN-LSTM based algorithms to improve Speech Emotion Recognition (SER) accuracy. We got an accuracy of 94.26 % which is quite good. The results of the present study have shown the ability of CNN-LSTM to study the emotional characteristics of speech signals from their low-level expression irrespective of sex and language. In spite of our networks' incredible performance, prospects for more improvements still remain.
As we know that emotions are embedded in several modes such as audio, visual and language. For future work, using multi-modal data will help us in approaching the problem of speech recognition from emotions in everyday life situations.

REFERENCES

1. Demis Hassabis, Dharshan Kumaran, Christopher Summer_eld and Matthew Botvinick. Neuroscience-inspired arti_cial intelligence. Neuron, 95(2):245{258, 2017.
2. El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572-587.
3. Louis Ten Bosch. Emotions, speech and the asr framework. Speech Communication, 40(1-2):213{225, 2003.
4. Harold Schlosberg. Three dimensions of emotion. Psychological review, 61 (2):81, 1954.
5. Thomas S Polzin and Alex Waibel. Detecting emotions in speech. In Proceedings of the CMC, volume 16. Citeseer, 1998.
6. Chul Min Lee and Shrikanth S Narayanan. Toward detecting emotions in spoken dialogs. IEEE transactions on speech and audio processing, 13(2):293{303, 2005
7. Matilda S. Emotion recognition: A survey. International Journal of Advanced Computer Research. 2015;3(1):14-19
8. Koolagudi SG, Rao KS. Emotion recognition from speech: A review. International Journal of Speech Technology. 2012;15(2):99-117.
9. Ayadi, E., Moataz, Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition features, classification schemes and databases. *Pattern Recognition*, *44*, 572–587.
10. Teager, H. (1990). Some observations on oral air flow during phonation. *IEEE Trans. Acoust. Speech Signal Process*, *28*, 599–601.
11.] Ren, Minjie, et al. "Multi-modal Correlated Network for emotion recognition in speech." *Visual Informatics* 3.3 (2019): 150-155.
12. G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. 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. Shanghai, China: IEEE, Mar. 2016, pp. 5200–5204.

13. A. Metallinou, M.Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, “Context-sensitive learning for enhanced audiovisual emotion classification,” IEEE Transactions on Affective Computing, vol. 3, no. 2, pp. 184–198, Jan. 2012.
14. T. Zhang,W. Zheng, Z. Cui, Y. Zong, and Y. Li, “Spatial-temporal recurrent neural network for emotion recognition,” IEEE Transactions on Cybernetics, vol. PP, pp. 1–9, Jan. 2018.
15. Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in Proceedings of the ACM International Conference on Multimedia, 2014, pp. 801-804.
16. T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, “Learning the speech front-end with raw waveform cldnns,” in Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, Sep. 2015, pp. 1–5.
17. G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network, Proc. of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China. pp. 5200-5204, 2016.
18. X. Li and X. Wu, “Long short-term memory based convolutional recurrent neural networks for large vocabulary speech recognition,” arXiv preprint arXiv:1610.03165, Sep. 2016.
19. A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, ``Speech emotion recognition from spectrograms with deep convolutional neural network," in Proc. IEEE Int. Conf. Platform Technol. Service (PlatCon), Feb. 2017, pp. 1-5.
20. Mirsamadi, S., Barsoum, E., Zhang, C., 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In: Proc. ICASSP, pp. 2227–2231.
21. Han, W., Ruan, H., Chen, X., Wang, Z., Li, H., Schuller, B., 2018. Towards temporal modelling of categorical speech emotion recognition. In: Proc. INTERSPEECH, pp. 932–936.
22. Niu, Yafeng, Zou, Dongsheng, Yadong, Niu, He, Zhongshi, Tan, Hua, 2017. A breakthrough in speech emotion recognition using deep retinal convolution neural networks. eprint arXiv:1707.09917.
23. Zhang, Shiqing, Xiaoming Zhao, and Qi Tian. "Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM." IEEE Transactions on Affective Computing (2019).
24. Beauregard, Gerald T., Mithila Harish, and Lonce Wyse. "Single pass spectrogram inversion." 2015 IEEE International Conference on Digital Signal Processing (DSP). IEEE, 2015.
25. Yenigalla, Promod, et al. "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding." Interspeech. 2018.
26. Stolar, Melissa N., et al. "Real time speech emotion recognition using RGB image classification and transfer learning." 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE, 2017.
27. Dennis, Jonathan, Huy Dat Tran, and Haizhou Li. "Spectrogram image feature for sound event classification in mismatched conditions." IEEE signal processing letters 18.2 (2010): 130-133.
28. Mao, Qirong, et al. "Learning salient features for speech emotion recognition using convolutional neural networks." IEEE transactions on multimedia 16.8 (2014): 2203-2213.
29. Lim, Wootaek, Daeyoung Jang, and Taejin Lee. "Speech emotion recognition using convolutional and recurrent neural networks." 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE, 2016.
30. Hajarolasvadi, Noushin, and Hasan Demirel. "3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms." Entropy 21.5 (2019): 479.
31. Peng, Zhichao, et al. "Speech emotion recognition using multichannel parallel convolutional recurrent neural networks based on Gammatone Auditory Filterbank." 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2017.
32. Zhang, Ruikai, et al. "Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain." IEEE journal of biomedical and health informatics 21.1 (2016): 41-47.
33. Zheng, W. Q., J. S. Yu, and Y. X. Zou. "An experimental study of speech emotion recognition based on deep convolutional neural networks." *2015 international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2015.
34. Neumann, Michael, 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).
35. Iliou, Theodoros, and Christos-Nikolaos Anagnostopoulos. "Classification on speech emotion recognition-a comparative study." animation 4 (2010): 5.
36. S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, Recurrent neural networks for emotion recognition in video, Proc. of ACM on International Conference on Multimodal Interaction (ICMI). pp. 467-474, 2015.
37. S. Zhang, S. Zhang, T. Huang, and W. Gao, Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching, IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576-1590, 2018.
38. P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, End-to-End Multimodal Emotion Recognition using Deep Neural Networks, IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, 2017.
39. E. Cambria, S. Poria, R. Bajpai, and B. Schuller, SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives, Proc. of COLING-2016, Osaka, Japan. pp. 2666-2677, 2016.
40. Chen, Shizhe, and Qin Jin. "Multi-modal dimensional emotion recognition using recurrent neural networks." Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. 2015.
41. Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Dobaie, A.M., 2018. Facial expression recognition via learning deep sparse autoencoders. Neurocomputing 273, 643–649.
42. Vielzeuf, V., Pateux, S., Jurie, F., 2017. Temporal multimodal fusion for video emotion classification in the wild. In: ICMI. ACM, pp. 569–576.
43. Zhao, Jianfeng, Xia Mao, and Lijiang Chen. "Speech emotion recognition using deep 1D & 2D CNN LSTM networks." Biomedical Signal Processing and Control 47 (2019): 312-323.
44. "Surrey Audio-Visual Expressed Emotion (SAVEE) Database". Researchgate, 2020, https://www.researchgate.net/publication/260311132_Surrey_AudioVisual_Expressed_Emotion_SAVEE_database.

AUTHORS PROFILE

Hafsa Qazi has done B.Tech degree in information technology from Baba Ghulam Shah Badshah University, Rajouri, J&K. She is persuing M.Tech degree in computer science from Shri Mata Vaishno Devi University, Katra, J&K. She has published a review paper in scopus indexed journal. Her area of interest is Machine Learning, and Deep Learning.

Baij Nath Kaushik received B.E. in Computer Science and Engineering from Nagpur University, Nagpur in 1997, Master of Technology from University School of Information Technology, GGSIPU, New Delhi in 2009 and Ph.D. in Computer Science from IIT Dhanbad, Dhanbad in 2016. He has more than 21 years of teaching and research experience and published many research papers in international journals and conferences of high repute. Presently, he is an Associate Professor in the School of Computer Science & Engineering, SMVDU, Katra, J&K. His research areas of interest include Machine Learning, Deep Learning, Nature Inspired Algorithms, Soft Computing and Parallel Algorithms.