A Self-Attention Based Hybrid CNN-LSTM for Speaker-Independent Speech Emotion Recognition

Paritosh Bhushan[1], Md Shah Fahad[2,*], Sanjay Agrawal[1], Paritosh Tripathi[1], Praveen Mishra[1], and Akshay Deepak[1]

ARTICLE INFO

Article history:
Received: 23 June 2022
Revised: 03 August 2022
Accepted: 28 August 2022

Keywords:
CNN-LSTM
Emotion recognition
Gender Embedding
Self-attention
Speaker-independent

ABSTRACT

Emotion recognition through speech signals has grown a lot as it facilitates human-machine interaction. In this manuscript, we propose a new self-attention-based convolution bidirectional long short term memory (Bi LSTM) with a gender embedding method. It beats the attention-based convolution BiLSTM method by 3.43%. Emotional characteristics are vulnerable to non-emotional influences such as the speaker, speech modes, and atmosphere. We have presumed that computing delta-deltas and deltas for individualized characteristics not retains only useful emotional intimation but also reduces the influence of irrelevant emotional variables, resulting in fewer miss classifications [1]. We have also presumed that the gender of the speaker also plays a significant role when we are dealing with speech signals. Therefore, we have done gender embedding and fed it as an input to our self-attention-based convolution BiLSTM along with extracted features. The self-attention system has shown exceptional abilities in learning task-specific feature representations. We have carried out the analysis of the interactive emotional dyadic motion capture (IEMOCAP) dataset that confirms the efficiency of the introduced method. We have achieved confirmed outcomes in terms of unweighted average recall (UAR).

1. INTRODUCTION

Emotion is a deliberate cerebral response (such as rage or fear) that is felt as a strong sensation directed towards a specific entity and is normally accompanied by physiological and behavioural changes in the body.One of the hottest fields of research in Machine learning is emotion recognition. It is a way of understanding human emotion. It has a wide range of possible uses, including robot interfaces, banks, call centers, video games, and so on. Knowledge about students' emotional states may help concentrate classroom orchestration or E-learning on improving teaching efficiency. It can also be helpful in areas such as pervasive computing to help support individuals. The ability to verbally identify anxiety may make providing help simpler.There are various ways of emotion recognition like emotion recognition through audio, video, text, or image. In the case of text data, many researchers work at the sentence level to isolate catchwords that reflect feelings in written documents. Emotion recognition in video is based on a combination of audio, picture, and text data. Emotion recognition in audio differs from emotion recognition in text in that it employs verbal sounds to derive emotions from audio. Speech is distinct. The speakers can't mask their feeling in their speech. As a result, the speaker's mood can be discerned by the questioner, who is a participant in the dialogue. Voiced speech comprises a wealth of material, including the speaker's emotional hints. When the orators hesitate between lines, speech includes not only the spoken speaker's speech but also silence and noise. In this paper, our focus is on emotion recognition through speech signals, often called Speech emotion recognition (SER). The role of SER is to diagnose human emotion from speech signals. It's an algorithm that uses tone and pitch to detect latent emotions. We would be able to predict emotions such as sadness, anxiety, neutral, regret, and many others using SER.

In conventional methods, the extraction of features was done manually. There are three stages in the overall design of SER using the conventional method. First of all, a speech recognition device derives features like pitch and energy. Then, with the aid of a feature extractor, these numbers are condensed into a reduced set of features. After this, a classifier is used to connect attributes with emotions using the data. Traditional classification strategies [2] require discriminating features to recognize speech emotion. Spectrum features like as mel frequency cepstral coefficients (MFCC), linear frequency cepstral coefficients

[1]*Institute of Engineering And Technology, Dr. Ram Manohar Lohia Avadh University,Ayodhya U.p., India.*
[2]*Birla Institute of Technology Mesra, India (835315).*
**Corresponding author:* Md Shah Fahad; Email: fahad8siddiqui@gmail.com.

(LFCC), and paralinguistic features such as F0, for example, can be used [3]. Kandali et al. [4] have proposed an emotion detection approach focused on the Gaussian-mixture model classifier and MFCC as features. Chenchah et al.[5] have proposed a framework for emotion detection using MFCC and LFCC is based on a support vector machine and the hidden Markov model. Using auto-associative neural networks help vector machines, and radial-basis function neural networks, the authors suggested to incorporated features with MFCC and the residual-phase feature for music-emotion recognition in Reference [6].

Conventional methods were both tiring and less accurate. Therefore, a new technique - feature extraction through artificial Network (DNN) was proposed which revolutionized the world of SER. In very less time, several models of DNN have been proposed. Some of them are a generalized DNN-based discriminant analysis approach to learning low-dimensional discriminative features designed for easy classification from a wide collection of acoustic features for emotion detection in Reference [7]. They outperform the SVM by a wide margin. The writers of Reference [8] have used convolutional neural networks (CNN) to face recognition and CNNs for face and voice recognition to identify the emotion of a recording. They proposed a powerful paradigm for facial expression emotion detection that outperformed the competition. The authors of Reference [9] suggested a deep dual recurrent encoder model uses both the text and audio signals to gain a deeper understanding of speech data.

In terms of obtaining distinct properties for SER, DNNs have demonstrated exceptional achievements. However, it appears that researchers are omitting the fact that DNN also uses individual highlights as input, which may be influenced by various speaking styles, speech quality, and environmental factors. [1] Personalized features are speech emotion features that particularly describe individual emotional information, reflect the speaker’s attributes, and cannot include universal emotional information which is unalterable across speakers, contents, and environments. The gender of a person also plays a vital role when we are trying to detect emotion through speech signals. We know that males have low pitch voices and females have high pitch voicesWe can't ignore the difference in pitch between male and female voices, which can have a significant impact.Therefore, to lessen the numerical variation of individual characteristics for various speakers, talking styles, and gender of the speaker, we have extracted both the features and gender information. Our proposed model concentrates on all the aspects of emotion recognition through speech. The significant contributions of this report are summed as follows:

We have calculated delta and delta-deltas for the log Mel-spectrogram and have done gender embedding. The mathematical differences for emotion unimportant factors can be effectively reduced by this.

We introduce a convolution BiLSTM for SER that better captures the log-Mels time-frequency relationship and leads to stable SER results.

In model 1, we used the attention model in addition to convolution BiLSTM and obtained an accuracy of 59.43%. We used the self-attention model in addition to convolution BiLSTM in our proposed model 2 and achieved an accuracy of 62.86%. This distinction is due to the fact that traditional attention was combined with recurrent neural networks to improve the model's efficiency (RNN). However, RNN is not used in the case of self-attention, and it performs much better and much faster. While the attention mechanism encourages output to focus on input while processing output, the self-attention model allows inputs to communicate with one another (i.e. measure attention of all other inputs with respect to one input). Therefore, it better representation of the proposed Self-Attention Based CNN-BiLSTM SER with Gender Information architecture.

The architecture of the proposed model consists eight parts:

(1) As CNN-BILSTM input, Log-Mel spectrogram (including static, delta, and delta-deltas) are retrieved from the speech signal's data.

(2) CNN is executed for local immutable feature selection with Log-Mel Spectrogram.

(3) Bi -directional recurrent neural networks are employed to learn volatile dependency among distinct local time-stamp unchanged properties.

(4) By tracing the emotional-relevant portions of the CNN-BiLSTM characteristics, a self-attention layer works to produce utterance-level characteristics.

(5) Gender label embedding to take the gender information into account.

(6) concatenate output from phases (4) and (5) to combine both the features.

(7) In order to get higher-level properties representations for improved assortment, utterance-level attributes are assigned to a fully connected layer.

(8) For the final assortment, high-level characteristics are combined into a softmax layer. Captures the log-Mels time-frequency relationship and leads to stable SER results.

Experimental outcomes show that the outlined process beats the baseline process for the IEMOCAP dataset.

The remaining section of the paper is organized as follows: The methodology is described in Section 2. Experimental design and outcomes are described in Section 3. The conclusion is described in Section 4.

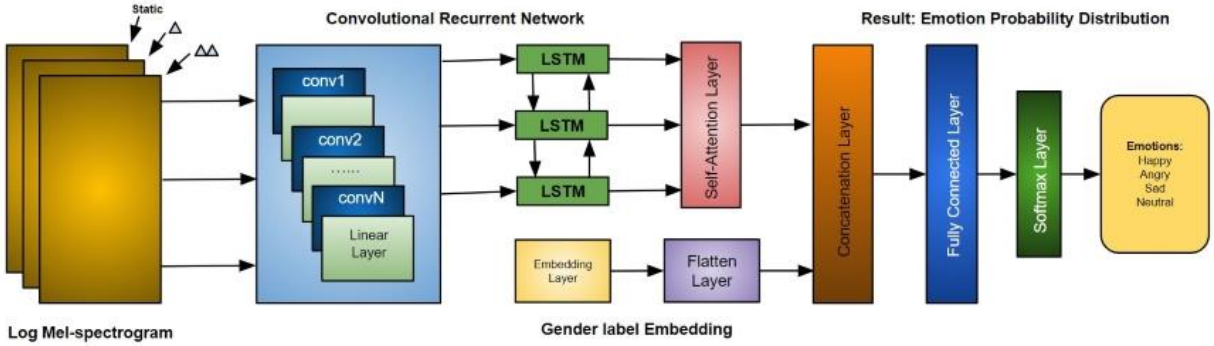

Fig. 1: Architecture for the proposed model

2. PROPOSED METHODOLOGY AND MODEL ARCHITECTURE

2.1. Sample preparation

The proposed architecture is a dual-input based DNN designed to benefit from two different but highly correlated information sources, namely

(i) speaker speech signals and (ii) speaker gender, for automated emotion prediction task. Both the inputs are processed separately via different neural network layers before being combined to produce a more suitable intermediate representation for emotion prediction.

The proposed architecture has major modules as follows:

1. Speech signal processing module
2. Gender processing module
3. Classification module

2.2 Speech signal processing module

The purpose of this module is to process and transform the raw speech signals into a viable intermediate representation. The steps included withinthis module are as follows:

2.2.1 3-D Log-Mels generation

The SER performance varies a lot depending on the speaker and theirspeaking style. In this paper, we employ log-Mels including delta and delta-deltas as the CNN contribution to solve this issue, where delta-deltas and deltas indicate the emotional transform mechanism. To minimise the difference between speakers, unit changes and zero-mean are applied to a speech signal, and the signal is divided into small edges accompanied with Hamming windows of 25ms and 10ms shift. By using Discrete Fourier Transform (DFT) calculated the power spectrum for every frame. Via going past the power spectrum using the Mel-filter bank i, output p_i is generated, in the form of mentioned in [1].

The log-Mels m_i is then obtained fom using the logarithm for p_i as shown in (1).

$$m_i = \log(p_i) \tag{1}$$

The formula given by (2) and (3) is used to figure out the deltas features m_i^{di} and m_i^{dd} of the log-Mels respectively.

$$m_i^d = \frac{\sum_{n=1}^{N} n(m_{i+n} - m_{i-n})}{2\sum_{n=1}^{N} n^2} \tag{2}$$

$$m_i^{dd} = \frac{\sum_{n=1}^{N} n\left(m_{i+n}^d - m_{i-n}^d\right)}{2\sum_{n=1}^{N} n^2} \tag{3}$$

After computation, the log-Mel spectrogram including delta and delta-deltas, we can receive a 3-dimentional characteristic representation $X \in R^{t \times f \times c}$ as the input of CNN, where the time (frame) t indicates length and f indicates the number of Mel-filter bank, and c indicates the number of feature channels. In this process, we assign f value as 40, the same as speech recognition [10], and c value assigs as 3,the representation done according to the order of the static, deltas, and delta-deltas.

2.2.2 CNN-BiLSTM model

- **CNN**: CNN is frequently used for deep learning, which avoids the requirement for custom extraction properties. CNN-based systems have been shown to achieve comparable to, or even better, exactness on SER tasks than classical frameworks **[11**-12], [13-14]. Convolution, pooling, and activation layers are the foundations of CNN. The number of information channels, the number of result capability maps, kernel size, and step all influence the convolutional layer.

CNN has demonstrated the tremendous success in the field of SER in recent years as in [14], [16-17]. With limited data, two-dimensional convolution outperforms One-dimensional convolution, according to William Chan and Ian Lane [18], and time region convolution is just as significant as frequency region convolution.

- **BiLSTM**: The concept of utilizing an LSTM network stems from the fact that can maintain long-term memory in a short period of time [19 Humans do not constantly begin their thoughts from scratch. When we read something, we interpret each word based on how we perceived the words before it. We don't discard anything and re-create everything from scratch. Our beliefs are steadfast [20]. We can improve

network performance by adopting BiLSTM and accommodate knowledge from the past to the present as well as from the present to the future. Most of the time, bidirectional LSTMs perform better than unidirectional LSTMs [21], [12].

CNN-BiLSTM is utilized to separate high-level characteristics for SER given a three-dimensional log-Mels. CNN-LSTM is composed of many 3-dimentional convolutional layers, in which one linear layer, one 3-D max-pooling layer, and one LSTM layer. Each convolutional layer filter size is 5 3, and the first convolutional layer, in particular includes 128 feature maps. The other rest of convolutional layers have 256 feature maps. We use max-pooling, with a pooling size of 2 2, and only after the first convolutional layer. Before passing three-dimensional CNN highlights into the LSTM layer a direct layer by adding, model parameters can be effectively reduced while maintaining accuracy. Thus, after the 3-D CNN, we include a linear layer with 768 output units for dimension reduction.

After completing three-dimensional CNN, we process the three-dimensional CNN order highlights in a bi-directional repeating neural network with long and transient memory cells for volatile summarizing, with 128 cells in each direction, to acquire an order of high-level feature representations in 256 dimensions.

- Self-Attention:

When we think of the term "attention" in English, we understand that it means to concentrate on something and pay closer attention. The foundation of Deep Learning's attention mechanism is the idea of directing the focus, and when processing the data, it gives particular features greater attention. This is a technique for reformulating the term representation based on the learned correlations with all terms in the sequential data. Generally, two types of attention mechanisms are used, *General-Attention* and *Self-Attention*. The interdependence in General-Attention is calculated between input and output elements, whereas the interdependence in Self-Attention is calculated within the input elements.

In SER, we have a one-word output representing the emotion class and sequential input in the form of a speech signal. We don't have an output sequence because it's a classification problem. As previously stated, General-Attention mechanism calculates interdependence using both the input and output sequences. However, in this case, the output sequence BiLSTM is just one word, which may not be capable of calculating as much significant attention as a sequence. As a result, we decided to use Self-Attention instead of General-Attention, which only calculates attention within the input sequence. So we decided to use *Self-Attention* in place of *General-Attention* which calculates attention within the input sequence only.

To compute self-attention, we must generate three new vectors from each of the encoder's input vectors. So we create a Query vector, a Value vector, and a Key vector for each word (here word is used for the vector, representing a small segment of the input speech segment of 3 seconds, as described in section 3). These vectors are generated by multiplying the embedding by three trained matrices created during the training cycle.

$$Qi = W_q xi \tag{4}$$

$$Ki = W_k xi \tag{5}$$

$$Vi = W_v xi \tag{6}$$

From these three vectors, the self-attention matrices are calculated as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

Since not all structure level CNN-BiLSTM characteristics contribute uniformly in order to show speech emotion, a Self-Attention layer has been used with a series of supreme-level representatives to concentrate on emotion pertinent section and generate discriminative utterance-level expressions to SER. Instead of simply conducting a mean/max pooling over time, we utilize the Self-Attention model to score the significance or a series of top-level expressions for last utterance-level emotion delineation. We calculate Self-Attention by passing it from a Self-Attention layer defined in *keras self-attention* with activation as 'relu' and the attention width as 15 and then perform *Global-Average-Pooling*.

2.3 Gender processing module

Every voice signal possesses some gender-specific characteristics. Speech recognition is significantly influenced by the speaker's gender, and research has revealed that some SER input features, including pitch, have different mean values for each gender. The effect of gender information on speech emotion recognition performance has been shown in the several research articles [22–23]. A gender-based model [24] has been demonstrated to be more accurate than a unified model for the same gender.

This module is intended to transform gender information into adense representation that can be efficiently combined with speech signal representations. This is composed of the following neural network layers:

2.3.1 Embedding Layer

So, in this proposed model we provide gender information as an input along with Log-Mels features.

We use *keras embedding layer* for embedding gender information and decide 300 as output dimensions.

2.3.2 Flatten Layer

The output from the embedding layer is next passed to a fully connected neural network layer (with 64 neuron units) and it is then flattened to get a 64-dimensional vector representation of gender information.

2.4 Classification module

Then, using the concatenate operation, we combine the output of the self-attention layer, which contains the speaker's utterance-level emotion features, and the 64-dimensional vector representation, which contains the speaker's gender information, to create a combined intermediate representation. The output layer, which has four neurons that represent four emotions, comes after the completely connected layers, which receive the combined intermediate representation (angry, sad, happy, neutral).

3. EXPERIMENTS

3.1 Experimental setup

To comply introduced model, we process independent-speaker SER probes. The IEMOCAP dataset has been used in this study to evaluate the performance of our suggested model. There are five sessions total, and a mixed gender group of speakers gives each one in both scripted and spontaneous settings. It has a total of 10039 utterances with a 16 kHz sampling rate, each lasting on an average 4.5 seconds.

We just consider the improvised data from the IEMOCAP dataset with four emotional categories: *happy*, *angry*, *sad*, and *neutral*. Since there are ten speakers in this database, we use a cross-validation of the 10-fold strategy in our assessments. Eight speakers are chosen as training data for each assessment, one speaker is chosen as validation data, and the remaining speaker is used as test data. By adjusting the parameter initialization, we can obtain a wide range of outcomes. We repeat each assessment five times with different random seeds to produce more precise results, and we then report the mean and standard deviation. Due to the skewed distributions of the test classes, we measure UAR on the test set. It should be noticed that all prototype structures, inclusive, and the number of ages, are picked through expanding the UAR of validate set.

We split the speech signal into equal-length 3 second chunks for improved parallel acceleration, and zero-padding is applied to utterances lasting up to 3 seconds. Each sub-portion predicts one emotion during preparation, and during testing, we evaluate the prediction of the full sentence by applying max pooling to the back probabilities of all sub-sentences. For each segment of the training data, we also generate a separate vector to contain gender labels. We label it in accordance with how the accompanying audio file is labelled. We assign 1 for males and 0 for females.

The preparation and test log-Mels are standardised using the global-mean and standard deviation of the training set, and the log-Mels are separated using the open EAR tool with a window size of 25 ms and a 10 ms shift. The Tensor Flow and Keras APIs and toolkits are used to implement the ACRNN architecture, and the model's parameters were simplified in accordance with the ability to minimise cross-entropy using a small collection of 40 instances while utilising Adam optimizer with Nestorov momentum. The speed is set at 0.9 and the rating for beginning the learning is set to 10^{-4}.

3.2 Baselines

We collate our proposed approach with different baselines:

1. Deep Neural Networks- Extreme Learning Machine [24] (DNN-ELM) is formed of 3-hidden layers with hidden units of 256, according to [25]. Statistical functions applying to the segment-level possibility yield utterance-level features, which are then fed into an ELM for the final decision.

2. CNN+LSTM with data augmentation [26]: In [26], 1-6 CNN layers and 1-4 Bi-LSTM layers are used along with the data augmentation. The stochastic gradient descent method is used for the optimization procedure.

Table 1. Accuracy for each fold using the proposed architecture

Fold	Session	Gender	WA (%)	UWA (%)	WA (%)	UWA (%)
			Model 1		Model 2	
1	1	F	73.17	61.90	62.00	66.40
2	1	M	52.19	63.50	71.70	67.11
3	2	F	54.80	71.50	71.10	68.00
4	2	M	62.01	74.80	61.50	74.93
5	3	F	69.69	58.13	64.20	62.32
6	3	M	66.66	57.46	61.30	70.00
7	4	F	64.70	64.30	68.00	64.14
8	4	M	69.23	64.92	69.23	65.00
9	5	F	58.82	34.82	58.20	42.34
10	5	M	57.91	43.00	55.98	48.40
10-fold Cross-Validation			**62.92**	**59.43**	**64.32**	**62.86**

Table 2: The Comparison of the proposed technique with the current works

Method	Accuracy
DNN-ELM	51.24 ± 7.24
CNN-LSTM	61.7
Model 1 (Proposed)	59.43 ± 11.57
Model 2 (Proposed)	**62.86** ± **9.43**

3.3 Model Comparison

In this paper we implemented two models:

1. Model 1: SER using CNN-LSTM with Attention mechanism and gender information.

2. Model 2: SER using CNN-LSTM with Self-Attention and gender information.

In Table 1 we have compared two models, Model1 and Model2. The main difference between these two methods is the applied attention mechanism. In model 1, we have used attention mechanism as in [26], in which CNN-LSTM works as the encoder whereas in Model 2, we implemented Self-Attention after the encoder part.

In both models, we used gender label embedding as another input and concatenate it with the output generated after the attention layer. For model evaluation, tenfold cross-validation technique is used. Model 2 outperforms the other three models in terms of UAR as mentioned in Table 1.

3.4 Experiment Results

We used weighted (WA) and unweighted (UA) accuracies to assess the model's accuracy. WA is the average accuracy calculated across the entire test set. The UA is a weighted average of the accuracies calculated separately for each emotion.

We calculate the metrics for each fold first and then average the results

Table 3: Confusion matrix using the proposed Model (1)

Emotion	Anger	Sad	Happy	Neutral
Anger	**69.40**	1.49	9.62	19.49
Sad	0.54	**69.90**	4.66	24.90
Happy	18.16	2.23	**30.41**	49.20
Neutral	10.60	9.42	12.08	**67.90**

Table 4: Confusion matrix using the proposed Model (2).

Emotion	Anger	Sad	Happy	Neutral
Anger	**67.55**	2.50	11.11	18.83
Sad	0.36	**84.08**	2.27	13.29
Happy	12.97	5.94	**39.50**	41.58
Neutral	5.96	19.18	14.88	**59.57**

Since UA is a more important feature for imbalanced datasets, we focused our efforts on achieving a high UA, as most other IEMOCAP papers have done.

Table 1 compares the baselines for UAR using our newly introduced method (Model 2). We began by contrasting our approach with the top-of-the-line DNN-ELM strategy discussed in [24]. Since the IEMOCAP dataset has a relatively uneven distribution of samples for each emotion class, and because we initially did not consider gender as a second input, many test samples were incorrectly classified as belonging to the "neutral" class. This is due to the fact that 48.2% of the samples from the imagined scenario fall inside the dataset's "neutral" category. We discovered after conducting numerous experiments that the predictions are not significantly biassed toward the "neutral" class when gender information is taken into account.Thus, we have changed the attention mechanism of model 2 and incorporate self-attention as a result of that 3.43 % improvement were seen in UAR.

Finally, in Tables 3 and 4, we present the confusion matrix for additional SER on IEMOCAP dataset analysis. Tables 3 and 4 show the confusion matrix for Model 1 and Model 2, respectively. These two confusion matrices show that Model 2 produces more precise results.

Here, in both cases, sad receives the greatest rate of acknowledgment while Happy receives the lowest.

4. CONCLUSION

In this article, authurs have suggested a technique for SER in which Self-Attention based CNN-BiLSTM is used along with gender information.

The process for extracting local correlations and knowledge about the global context from the raw audio signal, log-mel (static, deltas, delta-deltas) spectrograms is examined. Next, CNN and BiLSTM are used for high-level feature extraction. After this, the self-attention layer is used for focusing on emotion-related frames. Then we concatenate the gender information to it for SER.

Ten cross-fold validations for speaker-independent evaluations and four emotions (happy, sad, anger, and neutral) are applied to the IEMOCAP database to assess the performance of the suggested algorithm.

The introduction of gender information improves overall accuracy in our technique of SER by just 1% to 2%, but it significantly reduces the number of incorrect predictions made by our model. Due to uneven categories of emotions in the dataset, the majority of incorrect predictions before the gender information were in the "neutral" class. However, after taking into account the speaker's gender, we have come to the conclusion that gender aids our method of SER in learning its parameters more independently of the uneven distribution of training samples. In our method, self-attention performed better than general attention and produces superior outcomes, as was already mentioned.

The introduction of gender information improves overall accuracy in our technique of SER by just 1% to 2%, but it significantly reduces the number of incorrect predictions made by our model. Due to uneven categories of emotions in the dataset, the majority of incorrect predictions before the gender information were in the "neutral" class. However, after taking into account the speaker's gender, we have come to the conclusion that gender aids our method of SER in learning its parameters more independently of the uneven distribution of training samples. In our method, self-attention performed better than general attention and produces superior outcomes, as was already mentioned.

The findings of the experiment indicate that the self-Attention-based model (Model 2) performs better than model 1 and baselines in terms of overall UAR.

REFERENCES

[1] M. Chen, X. He, J. Yang, H. Zhang; 2018. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition; IEEE Signal Processing Letters (Volume: 25, Issue: 10): 1440 - 1444.

[2] Jianfeng Zhao, Xia Mao, Lijiang Chen; 2019; Speech emotion recognition using deep 1d and 2d cnn lstm networks, Biomed Signal Process Control (Volume: 47): 312–323.

[3] Laurence Vidrascu, Laurence Devillers: 2007; Five emotion classes detection in real-world call centre data: The use of various types of paralinguistic dWorkshop on Paralinguistic Speech Between Models and Data, Saarbr¨ucken, Germany, 2–3 August.

[4] C. Kandali, A. Routray, T. Basu;2008. Emotion recognition from assamese speeches using mfcc features and gmm classifier. Proceedings of the 2008 IEEE Region 10 Conference (TENCON 19-21 November , Hyderabad, India 1–5.

[5] F. Chenchah, Z. Lachiri; 2015. Acoustic Emotion Recognition Using Linear and Nonlinear Cepstral Coefficients: International Journal of Advanced Computer Science and Applications (Vol. 6, issue 11) : 135–138.

[6] N. Nalini, S. Palanivel; 2016. Music emotion recognition: The combined evidence of MFCC and residual phase: Egyptian Informatics Journal(Volume 17, Issue 1): 1–10.

[7] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, B. Schuller; 2011. Deep neural networks for acoustic emotion recognition: Raising the benchmarks, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) May 22-27, 5688–5691.

[8] S. Albanie, A. Nagrani, A. Vedaldi, A. Zisserman; 2018.Emotion recognition in speech using cross-modal transfer in the wild, Proceedings of the ACM Multimedia Conference on Multimedia Conference—MM ’18, Seoul, Korea 15 october: 292–301.

[9] S. Yoon, S. Byun, K. Jung; 2018. Multimodal speech emotion recognition using audio and text, Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 112–118.

[10] Keren, Gil, B. Schuller; 2016. Convolutional RNN: An enhanced model for extracting features from sequential data, International Joint Conference on Neural Networks (IJCNN) IEEE: 3412-3419.

[11] NGO Quang Uoc1, NGO Tri Duong , LE Anh Son , and BUI Dang Thanh; 2022. A Novel Automatic Detecting System for Cucumber Disease Based on the Convolution Neural Network Algorithm, GMSARN International Journal (Volume: 16): 295-301

[12] S. Mirsamadi, E. Barsoum, C. Zhang; 2017. Automatic speech emotion recognition using recurrent neural networks with local attention, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 2227-2231.

[13] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, S. Zafeiriou; 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network, Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference: 5200–5204.

[14] Z. Aldeneh, E. M. Provost; 2017. Using regional saliency for speech emotion recognition; Acoustics, Speech and Signal Processing (ICASSP): 2741–2745.

[15] S. Zhang, T. Huang, W. Gao; 2017. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Transactions on Multimedia (Volume: 20 Issue: 6): 1576 - 1590.

[16] Q. Mao, M. Dong, Z. Huang, Y. Zhan; 2014. Learning salient features for speech emotion recognition using convolutional neural networks; IEEE Transactions on Multimedia (Vol.:16, issue 8): 2203–2213.

[17] W. Chan, I. Lane; 2015. Deep convolutional neural networks for acoustic modeling in low resource languages, IEEE International Conference on Acoustics, Speech and Signal Processing, 19-24 April, Australia: South Brisbane, QLD, 2056 - 2060.

[18] S. Hochreiter, J. J. U. Schmidhuber; 1997. Long short-term memory, Neural Computation, (Volume :9, issue:8): 1735-1780.

[19] C. Olah, Understanding lstm networks, colahs blog (2015).

[20] Tripathi, Samarth, H. Beigi; 2018. Multi-modal emotion recognition on iemocap dataset using deep learning, arXiv preprint arXiv:1804.05788.

[21] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, A. Sciarrone; 2013. Gender-driven emotion recognition through speech signals for ambient intelligence applications, IEEE Trans. Emerging Topics in Computer (volume:1, issue:2): 244-257.

[22] R. Xia, J. Deng, B. Schuller, Y. Liu, 2014. Modeling gender information for emotion recognition using denoising auto encoder; IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP): 990-994.

[23] T.-W. SUN; 2020. End-to-end speech emotion recognition with gender information, Graduate Student Member, IEEE Access: (volume:8): 152423 - 152438.

[24] K. Han, D. Yu, I. Tashev; 2014. Speech emotion recognition using deep neural network and extreme learning machine, Proceedings of Interspeech:14-16 September: Singapore:223-227.

[25] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, B. Schmauch; 2018. Cnn+lstm architecture for speech emotion recognition with data augmentation, Proc. Workshop on Speech, Music and Mind :21-25.

[26] Bahdanau Dzmitry, K. Cho, Y. Bengio; 2015. Neural machine translation by jointly learning to align and translate., ICLR:1-15.