International Journal of Informatics, Information System and Computer Engineering

2-D Attention-Based Convolutional Recurrent Neural Network for Speech Emotion Recognition

Akalya Devi C, Karthika Renuka D, Aarshana E Winy, P C Kruthikkha, Ramya P, Soundarya S

Assistant Professor, 2UG Scholar, Department of Information Technology,PSG College of Technology, Coimbatore, India

*Corresponding Email: cad.it@psgtech.ac.in

ABSTRACTS

Recognizing speech emotions is a formidable challenge due to the complexity of emotions. The function of Speech Emotion Recognition (SER) is significantly impacted by the effects of emotional signals retrieved from speech. The majority of emotional traits, on the other hand, are sensitive to emotionally neutral elements like the speaker, speaking manner, and gender. In this work, we postulate that computing deltas for individual features maintain useful information which is mainly relevant to emotional traits while it minimizes the loss of emotionally irrelevant components, thus leading to fewer misclassifications. Additionally, Speech Emotion Recognition (SER) commonly experiences silent and emotionally unrelated frames. The proposed technique is quite good at picking up important feature representations for emotion relevant features. So here is a two dimensional convolutional recurrent neural network that is attention-based to learn distinguishing characteristics and predict the emotions. The Mel-spectrogram is used for feature extraction. The suggested technique is conducted on IEMOCAP dataset and it has better performance, with 68% accuracy value.

ARTICLE INFO

Article History:
Received 18 Dec 2022
Revised 20 Dec 2022
Accepted 25 Dec 2022
Available online 26 Dec 2022

Keywords:
2-D,
Attention-Based,
Convolutional Recurrent Neural Network,
Speech Emotion Recognition

1. INTRODUCRION

The significance of human speech emotion recognition has increased recently to increase the quality and efficiency of interactions between machines and humans (Khalil et al., 2019). Due to the difficulty in defining both natural and artificial emotions, recognizing human emotions is a challenging task all on its own. Extraction of the spectral and prosodic elements that would lead to the accurate assessment of emotions has been the subject of numerous investigations (Tzirakis et al., 2018).

Recognition of speech emotions is a technique that uses a processor to extract emotional information from speech signals (Chen et al., 2018). It then compares and analyzes the collected emotional information together with the distinctive factors. After the emotional information is extracted, various techniques and concepts are used to predict the emotions of speech signals (Khalil et al., 2019). Speech emotion detection is now a rapidly developing discipline bridging the interaction between robots and humans. It is also a popular study area in signal processing and pattern recognition.

Emotions are incredibly important to human mental health. It is a way of expressing one's thoughts or state of mind to others. The major objective of SER is to improve human-machine interaction (HMI). It can also be used with lie detectors to monitor a subject's psychophysical state (Lalitha et al., 2015). Onboard car driving systems, dialogue systems for spoken languages used in call center conversations and the utilization of speech emotion patterns in medical applications are a few instances of SER. HMI systems still have a lot of problems that need to be resolved, especially when they are shifted from being tested in labs to being used in actual operations. Therefore, efforts are required to effectively resolve all these problems and enhance machine emotion perception.

Recently, Deep Neural Networks (DNNs) have gained popularity and made revolutionary strides in a number of machines learning fields, including the continuous effect identification field. In most of the studies, hand-crafted features are used to feed the DNN architectures. Many DNN architectures have been put forth in that approach, including Convolutional Neural Networks (CNNs) and Long-Short Term Memory (LSTM) networks. Mao et al. (Mao et al., 2014) first used Convolutional neural networks (CNN) and demonstrated great scores on numerous benchmark datasets for learning affective-salient features for SER. Recurrent neural networks (RNNs) were used by Lee et al. (Lee & Tashey, 2015) to train SER on long-range temporal correlations. In order to train a convolutional recurrent neural network (CRNN) to predict continuous valence space, Trigeorgis et al. (Trigeorgis et al., 2016) directly used the raw audio data.

Additionally, structures connecting the output and the input segments have been learned with significant effectiveness using attention mechanism-based RNNs. RNNs based on attention mechanisms are ideally suited to the SER tasks. First, speech is basically a sequence of data with different lengths. The majority of speech signals annotate emotion labels at the utterance level even though utterances sometimes have lengthy pauses and frequently have a

short word count. Selecting emotional-relevant frames for SER is very crucial. In this paper, we extend our model to yield affective salient characteristics for the final emotion categorization using a CRNN-based attention mechanism.

In this study, we combine CRNN and an attention model to create a unique architecture for SER dubbed 2-D attention-based convolutional recurrent neural networks (ACRNN). The following is a summary of this paper's main contributions:

1) We suggest a unique 2-D CRNN for SER that enhances the ability to understand the time-frequency relationship.

2) We employ an additional attention model to automatically concentrate on the emotion-crucial frames and offer discriminative utterance-level characteristics for SER to cope with silence critical frames and emotion-irrelevant frames.

3) Experimental results contain athe ccuracy, recall, precision and confusion matrix of our proposed model.

It is well known that most speech emotion datasets only have utterance-level class labels. Most sentences, however, contain silent regions, short pauses, transitions between phonemes, unvoiced phonemes, and so on. It is clear that not all parts of a sentence are emotionally connected. Unfortunately, LSTM does not handle this situation well when analyzing acoustic characteristics extracted from voice. In the current study, emotion classification is useful for distinguishing between emotionally-relevant and emotionally-irrelevant frames. In emotion classification, it is useful to know whether the speech frame is voiced or unvoiced. Currently, there are two types of commonly used methods: manually extracting emotionally relevant speech frames and using models to learn how to distinguish automatically.

However, as manual extraction requires different thresholds on different data sets, it has some limitations in terms of feasibility. Human emotional expression is often gradual and thus each voiced frame is useful for emotion classification. Attention mechanisms can better match human emotional expression by capturing only the affective frames. Local attention was added to LSTM and different weights were assigned to each frame with varying emotional intensity.

2. LITERATURE SURVEY

On the IEMOCAP dataset, Sarthak Tripathi and Homayoon Beigi performed multimodal emotion detection and determined the best individual architectures for classification of each modality using data from speech, text and motion capture. The design of their merged model is modular. This makes it possible to upgrade any individual model without affecting the other modalities. They utilized motion captured data and 2D convolutions in place of video recordings and 3D convolutions (Tripathi et al., 2018).

For the Arabic dataset KSUEmotions, Mohammed Zakariah and Yaser Mohammad Seddiq performed Speech emotion Recognition. The feature extraction method made use of the time-frequency data from the spectrogram, as well as numerous modification and

filtering techniques were used. Although the system was tested at the file and segment levels, it was trained at the segment level (Maji & Swain, 2022).

To automatically extract affective salient features from raw spectral data, Yawei Mu and Luis A. Hernandez Gomez presented a distributed convolution neural network (CNN). From the CNN output, they then applied a bidirectional recurrent neural network (BRNN) to obtain temporal information. Finally, they used the attention mechanism to target the emotion-relevant portions of utterance in the BRNN output sequence (Jiang et al., 2021).

A Convolutional-Recurrent Neural Network with Multiple Attention Mechanisms (CRNN-MA) was proposed by P. Jiang, X. Xu, H. Tao, L. Zhao, and C. Zou for SER. It uses extracted Mel-spectrums and frame-level features in parallel Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) modules, respectively. A multi-dimensional attention layer and multiple self-attention layers in the CNN module on frame-level weight components (Yadav et al, 2021) are some of the strategies they established for the suggested CRNN-MA.

Yadav, O. P., Bastola, L. P., and Sharma, J. presented the Convolutional Recurrent Neural Network (CRNN), which combines Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM), to learn emotional features from log-mel scaled spectrograms of spoken utterances. Convolution kernels of CNN are used to learn local features and a layer of BiLSTM is chosen to learn the temporal dependencies from the learnt local features. Speech utterances are pre-processed to cut out distracting sounds and unnecessary information. Additionally, methods for increasing the number of data samples are researched, and the best methods are chosen to improve the model's recognition rate (Lim et al., 2016).

Without employing any conventional hand-crafted features, Wootaek Lim, Daeyoung Jang, and Taejin Lee developed a SER approach based on concatenated CNNs and RNNs. Particularly for computer vision tasks, Convolutional Neural Networks (CNNs) have exceptional recognition ability. Recurrent neural networks (RNNs) also perform sequential data processing tasks to a great extent with high degree of success. The classification result was proven to have higher accuracy than that attained using traditional classification methods by utilizing the proposed methods on an emotional speech database (Gayathri et a., 2020).

Silent frames and inappropriate emotional frames are frequent problems for Speech Emotion Recognition (SER). Meanwhile, the attention process has proved to be exceptionally effective at learning relevant feature representations for particular activities. Using the Mel-spectrogram with deltas and delta-deltas as input, Gayathri, P., Priya, P. G., Sravani, L., Johnson, S., and Sampath, V. presented a Convolutional Recurrent Neural Networks (ACRNN) based on attention to learn discriminative features for SER. Finally, test results

demonstrated the viability of the suggested approach and achieved cutting-edge performance in terms of unweighted average recall (Gayathri et a., 2020).

2.1. Proposed Models and Experimental setup

A convolutional recurrent neural network with a 2D attention base, serves as the proposed model for speech emotion recognition.

2.2. Speech Emotion Recognition

This section explains the proposed 2D attention based convolutional recurrent neural network. Convolutional neural network, or CNN or ConvNet, is particularly adept at processing input with a grid-like architecture, like an image. A binary representation of visual data is a digital image. Recurrent neural networks (RNNs) are a type of neural network in which the results of one step are fed into the next step's computations. It employs sequential data or time series data. The Convolutional Recurrent Neural Network (CRNN) model uses the outputs and hidden states of the recurrent units in each frame to extract features from the successive windows by feeding each window frame by frame into the recurrent layer. Here we combine an attention mechanism together with CNN and RNN that enables easier and higher-quality learning by concentrating on certain portions of the input sequence inorder to predict a particular portion of the output sequence.

Feature extraction is a process that converts raw data into manageable numerical features while preserving the original data's information. Feature extraction when compared to using machine learning or deep learning models on the raw data directly, produces better outcomes. For the feature extraction Log Mel-spectrogram is used. The ACRNN architecture, which combines CRNN with an attention model, is used. Then, as depicted in Fig. 1, a fully linked layer and a softmax layer for SER are introduced.

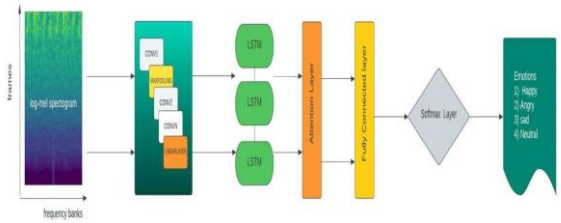

Fig. 1. ACRNN architecture

CNN has recently demonstrated remarkable accomplishments in the SER field. The time domain and frequency domain are equally important and 2-dimensional convolution performs better with less data than 1-dimensional convolution. The SER findings, however, vary greatly between speakers because of huge variation in tone, voice and other unique characteristics. The log-Mels with deltas and delta-deltas act as the ACRNN input to handle this variation, where the deltas and delta-deltas describe the emotional transformation process.

The mel scale has a range of pitches that to the human ear, appear to be equally distant from one another. The distance in hertz between mel scale values, often known as "mels," increases as the frequency increases. Mel, which stands for melody, denotes that the scale is founded on pitch comparisons.

Extensive tests have shown us that the Mel spectrum is better compatible with the human auditory sense characteristic, which exhibits the linear distribution under 1000 Hz and the logarithm growth above 1000 Hz and hence this point is used to obtain the Log-Mel spectrum static. The link between the frequency and the Mel spectrum is interrelated.

A mel spectrogram renders frequencies over a specific threshold logarithmically (the corner frequency). For instance, in the spectrogram with a linear scale, the vertical space between 1,000 and 2,000 Hz is half that between 2,000 and 4,000 Hz. The distance between the ranges is almost the same in the mel spectrogram. Similar to how we hear, this scaling makes similar low frequency sounds simpler to identify from similar high frequency noises. A frequency-domain value is multiplied by a filter bank to create the output of a mel spectrogram.

When a speech signal is given with zero mean and unit variance, it is used to minimize the differences between speakers. The signal is then divided into small frames using Hamming windows with a shift of 10 ms and a 25 ms duration. The power spectrum is then placed through the Mel-filter bank I to produce output pi, and the output is then used to calculate the power spectrum for each frame using the Discrete Fourier transform (DFT) i. The logarithm of pi is then used to produce the log-Mels mi, as shown by (1). To determine the log-Mels' deltas features, we use the following formula (2). N is often selected as (2). Similarly, the delta-deltas features are calculated using the time derivative of the deltas, as seen in (3).

Generate a 3-D feature representation for the CNN input X Σ 〖 〖 R〗 ^(t×f×c)〗 ^ by t stands for the time (frame) length, f for the number of Mel-filter banks, and c for the number of feature channels when computing the log-Mels with deltas and delta-deltas. As in speech recognition [17], we set f in this task to 40 and c to 3, which stand for static, deltas, and delta-deltas, respectively.

$$m_i = log(p_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

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

$$m_i^{dd} = \frac{\Sigma \ \sum_{n=1}^{N} n(m_{i+n}^d - m_{i-n}^d)}{2\Sigma \ \sum_{n=1}^{N} n^2} \quad \ldots\ldots\ldots..(3)$$

2.3. ACRNN architecture:

In this part, we integrate CRNN with an attention model along with 2-D log-Mels. 2-D CNN is used to perform convolution in a patch that only contains a few frames on the entire log-Mels. The long short-term memory (LSTM) is then fed with 2-D CNN sequential characteristics for temporal summarization. A series of high-level features are then entered into the attention layer, which outputs utterance-level features. Finally, utterance-level characteristics are used as the fully connected layer input to obtain higher level features for SER.

1)CRNN model: High-level features for SER are retrieved using CRNN from given 2-D log-Mels. The CRNN used here

consists of several 2-D convolution layers, one 2-D max-pooling layer, one linear layer, and one LSTM layer. Each convolutional layer has a 5 x 2 filter size, with the first convolutional layer having 128 feature maps and the subsequent convolutional layers having 256 feature maps. After the first convolutional layer, we only use one max pooling layer and the pooling size is 2 x 2. The model parameters can be effectively reduced without compromising accuracy by adding a linear layer before feeding 2-D CNN features into the LSTM layer. As a result, we find that the linear layer with 768 output units is appropriate when added as a dimension-reduction layer after the 2-D CNN. We perform a 2-D CNN and then feed the 2-D CNN sequence features via a bidirectional RNN with 128 cells in each direction for temporal summarization. As a result, a sequence of 256-dimensional high-level feature representations are obtained.

2)Attention Layer: Due to the fact that not all frame-level CRNN features equally contribute to the representation of speech emotion, an attention layer is employed to focus on emotion-relevant sections and produce discriminative utterance-level representations for SER.

Instead of only using a mean/max pooling across time, the significance of a number of high-level representations to the utterance-level emotion representations is rated using an attention model.

In particular, first determine the normalized weight using a softmax function and the LSTM output ht at time step t. Then, as illustrated, perform a weighted sum on ht using the weights to determine the utterance-level representations (5).

Finally, feed the utterance-level representations through a fully connected layer with 64 output units to obtain higher level representations that help the softmax classifier map the utterance representations into N different spaces, where N is the number of emotion classes. The fully connected layer is subjected to batch normalization (Gayathri et a., 2020) to expedite training and enhance generalization performance.

$$a_t = \frac{exp(W.ht)}{\Sigma \ \ _{T=1}^{T} exp(W.ht)} \quad \ldots\ldots\ldots(4)$$

$$c = \Sigma \ \ _{t=1}^{T} a_t \, h_t \quad \ldots\ldots\ldots\ldots(5)$$

We conduct SER experiments using the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) to assess performance of our proposed model. There are five sessions of IEMOCAP each having utterances having duration on an average lasting for 4.5 seconds and the rate of each sample being 16 kilohertz. Every session here is presented by two speakers (a male and female) in both scripted scenes and improvised scenes. Only four emotions are considered here- angry, sad, happy and neutral. Cross validation used here for evaluation is 10-fold. Out of the total ten speakers, eight speakers are chosen for training the model, one speaker is chosen for testing and the other speaker is chosen for validation.

Consequently, we perform each evaluation multiple times using various random seeds in order to obtain more reliable findings. We divide the signal into 3 segments which are all equal in

length for improved acceleration which is parallel. We have also padded with zeros for speech utterances which are lasting less than 3 s. Training set's standard deviation and mean(global) are used for normalization of log-Mels of testing and training data, with 25 ms as the size of the window and a shift of 10 ms. Tensorflow and Keras libraries are installed for implementation (See Figs. 2-4).

	precision	recall	f1-score
ang	0.72	0.25	0.38
exc	0.63	0.66	0.64
neu	0.51	0.55	0.53
sad	0.54	0.82	0.65
accuracy			0.56
macro avg	0.60	0.57	0.55
weighted avg	0.59	0.56	0.54

Fig. 2. Workflow for Azure Machine Learning

	precision	recall	f1-score
ang	0.58	0.47	0.52
exc	0.74	0.56	0.64
neu	0.56	0.52	0.54
sad	0.53	0.79	0.63
accuracy			0.58
macro avg	0.60	0.59	0.58
weighted avg	0.59	0.58	0.58

Fig. 3. Workflow for Azure Machine Learning

	precision	recall	f1-score
0	0.72	0.72	0.72
1	0.71	0.70	0.70
2	0.46	0.29	0.35
3	0.72	0.82	0.77
accuracy			0.68
macro avg	0.65	0.63	0.64
weighted avg	0.67	0.68	0.67

Fig 4. Classification report of 2D attention based CRNN

Fig. 1 represents the classification report of 1D CNN LSTM which has an accuracy of 56%, precision value of 59% and recall value of 56%. Fig. 2 represents the classification report of Temporal 2-D CNN which has an accuracy of 58%, precision value of 59% and recall value of 58%. Fig. 3 represents the classification report of 2-D ACRNN which has an accuracy of 68%, precision value of 67% and recall value of 68%. Thus, our ACRNN model's performance is superior while compared with other models (See Fig. 5).

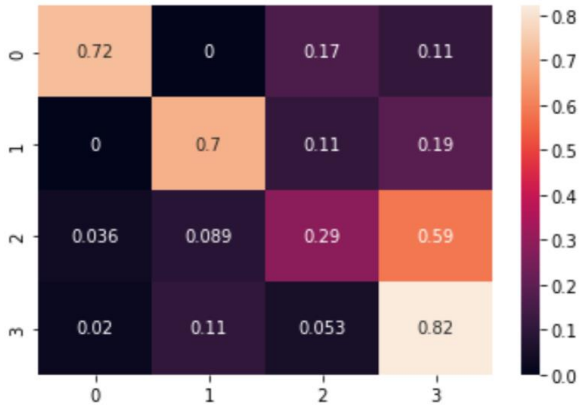

Fig 5. Classification report of 2D attention based CRNN

Fig. 4 displays the confusion matrix of the ACRNN model. There are four emotions- 0 represents angry, 1 represents sad, 2 represents happy and 3 represents neutral. The diagonal values represent the correctly predicted values. The accuracy of our proposed model 2-D CRNN is 68% which is higher than the accuracy of 1D CNN LSTM and T-2D CNN. Weighted precision of our model is 0.67, weighted recall of our model is 0.68. Weighted F1 score of our model is 0.67. All these values are higher than the corresponding values in 1D CNN LSTM and T-2D CNN. Thus, our model outperforms similar SER models with greater values for all metrics. 3-D

attention based CRNN implemented in the paper Chen (Chen et al., 2018) has average recall value of 64.74%. Our 2-D attention based CRNN has outperformed it with a recall value of 68% (See Fig. 6).

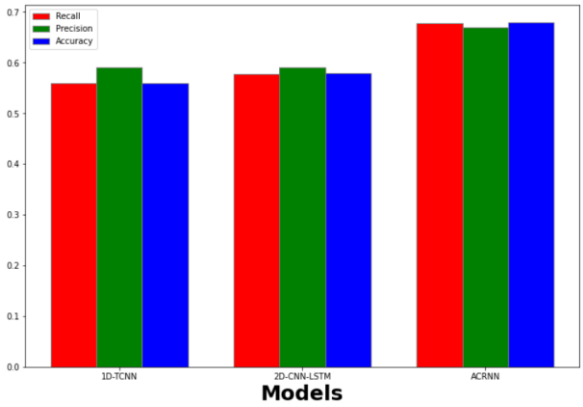

Fig 6. Comparison of Models and Their Evaluation Metrics

Fig. 5 shows the plot between models and their evaluation metrics. Our model comes out to be the best in all metrics while comparing with the other two models.

REFERENCES

Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, *25*(10), 1440-1444.

Gayathri, P., Priya, P. G., Sravani, L., Johnson, S., & Sampath, V. (2020). Convolutional Recurrent Neural Networks Based Speech Emotion Recognition. *Journal of Computational and Theoretical Nanoscience*, *17*(8), 3786-3789.

Huang, C. W., & Narayanan, S. S. (2016, September). Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition. In *Interspeech* (pp. 1387-1391).

Huang, C., Gong, W., Fu, W., & Feng, D. (2014). A research of speech emotion recognition based on deep belief network and SVM. *Mathematical Problems in Engineering*, *2014*.

Jiang, P., Xu, X., Tao, H., Zhao, L., & Zou, C. (2021). Convolutional-Recurrent Neural Networks with Multiple Attention Mechanisms for Speech Emotion Recognition. *IEEE Transactions on Cognitive and Developmental Systems*.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, *7*, 117327-117345.

Lalitha, S., Mudupu, A., Nandyala, B. V., & Munagala, R. (2015, December). Speech emotion recognition using DWT. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-4). IEEE.

Lee, J., & Tashev, I. (2015, September). High-level feature representation using recurrent neural network for speech emotion recognition. In *Interspeech 2015*.

Lim, W., Jang, D., & Lee, T. (2016, December). Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)* (pp. 1-4). IEEE.

Maji, B., & Swain, M. (2022). Advanced Fusion-Based Speech Emotion Recognition System Using a Dual-Attention Mechanism with Conv-Caps and Bi-GRU Features. *Electronics*, *11*(9), 1328.

Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE transactions on multimedia*, *16*(8), 2203-2213.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003, December). Detection of stress and emotion in speech using traditional and FFT based log energy features. In *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint* (Vol. 3, pp. 1619-1623). IEEE.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016, March). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5200-5204). IEEE.

Tripathi, S., Tripathi, S., & Beigi, H. (2018). Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*.

Tzirakis, P., Zhang, J., & Schuller, B. W. (2018, April). End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5089-5093). IEEE.

Yadav, O. P., Bastola, L. P., & Sharma, J. (2021). Speech Emotion Recognition using Convolutional Recurrent Neural Network.