Emotional Speech recognition

Rohan Jain, Joykirat Singh and **Sudarshan Buxy**

rohan19095@iiitd.ac.in
joykirat19166@iiitd.ac.in
sudarshan19279@iiitd.ac.in

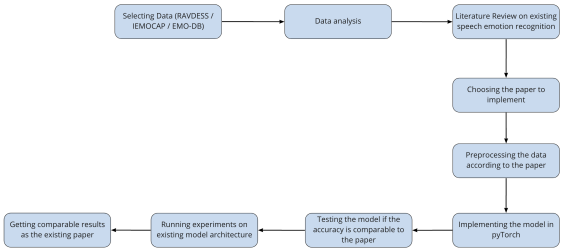
1 Introduction and Motivation

In recent years we have seen a sudden growth in virtual voice assistants. More people are interacting with Siri, Alexa, Cortana and Google Assistants. Interaction that ignore a user's emotional states or fails to detect the appropriate emotion and drastically decrease its performance. It can also make the assistant perceived as cold, untrustworthy. Therefore speech emotion recognition is becoming an increasingly relevant task.

2 Problem Statement

Deep Learning has been considered as an emerging research field in machine learning and has gained more attention in recent years. Application of deep learning such as SER (Speech emotion recognition) have several advantages over traditional methods, including their capability to detect the complex structure and features without the need for manual feature extraction and tuning.
Our objective to study, review and reproduce the results of the state of the art SER deep learning models.

3 Project pipeline



4 Literature Survey

4.1 Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions(Tripathi et al., 2019)

The author proposed a model that uses speech features and transcriptions(text) to predict emotions in speech. Different experiments were performed on speech transcriptions and speech features independently as well as together to achieve greater accuracies than existing state of the art model. IEMOCAP dataset was used in this study, and different models accuracies are shown in the table 1.
The text only based CNN model (Model 1) failed to capture all the low level features of speech signals. The combined MFCC and Spectrogram based CNN model achieves an overall emotion detection accuracy improvement close to 4% over existing state-of-the-art methods. The combined Text-MFCC model performs even better and beats the benchmark class accuracy by 5.5% and the overall accuracy by close to 7%.

Table 1: Comparison of accuracies

Methods	Input	Accuracy
Model 1	Text	64.4
Model 2A	Spectrogram	71.2
Model 2B	Spectrogram	71.3
Model 3	MFCC	71.6
Model 4A	Spectrogram + MFCC	73.6
Model 4B	Text + Spectrogram	75.1
Model 4C	Text + MFCC	76.1

4.2 Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings(Pepino et al., 2021)

This paper proposes a transfer learning method for speech emotion recognition where features are extracted from pre trained wav2vec 2.0 models.

Ouputs of several layers from the pretrained models were combined using trainable weights with the downstream model. Wav2Vec is a framework for self-supervised learning of representations from raw audio. Wav2Vec have 3 stages local encoder, contextual encoder, quantisation module. In this paper 2 different type of Wav2Vec model were use for finetuning.

- Wav2Vec2-FT : Model finetuned for ASR using a 960 hour subset of libriSpeech.
- Wav2Vec2-PT : Wav2Vec 2.0 base model pretrained in LibriSpeech without finetuning.

Table 2: Comparison of accuracies

Pretrain	Features	IEMOCAP	RAVDESS
None	eGeMAPS	52.4	57.0
	Spectrogram	49.8	44.5
W2VPT	Local enc.	60.3	65.4
	Cont. enc.	58.5	69.0
	All layers	67.2	84.3
W2VFT	Local enc.	57.3	58.8
	Cont. enc.	44.6	37.5
	All layers	63.8	68.7

Two systems based on eGeMAPS and spectrogram features where considered as baselines. wav2vec2-PT features perform better than wav2vec2-FT features in all cases. The model using only the contextualized encoder outputs for the wav2vec2-FT model has the worst results in the table. This maybe because when the model is finetuned for an ASR task, information that is not relevant for that task but might be relevant for speech emotion recognition is lost from the embeddings.

4.3 A Novel end-to-end Speech Emotion Recognition Network with Stacked Transformer Layers(Wang et al., 2021)

The mainstream pipeline in SER is to extract features and then aggregate them to form an emotional embedding. This paper introduces stacked multiple transformer layers in the aggregation part of the pipeline to improve results on the IEMOCAP dataset. The authors used 6 stacked transformer layers. Each layer had 2 sublayers: a self attention layer and one position wide feed forward network. In the proposed model, raw signal is preprocessed to extract Low level descriptors. LLDs are passed to a CNN-BiLSTM module to extract contextual representation which is feeded into stacked transformer layers to enhance the representation. This representation is the emotional embedding which is classified using a softmax layer. The paper used a subset of the IEMOCAP dataset with only 4 classes: Neutral, Happy, Sad and Angry. The researchers used 7.5s clips for training and testing. They used 5 fold cross validation to prevent overfitting. Weighted accuracy and unweighted accuracy were used as evaluation metrics. Weighted accuracy increased from 84% to 91.2% and unweighted accuracy increased from 82% to 92%. Neutral label was the toughest to predict. The future scope of the paper is to generalize the proposal and investigate its effectiveness on a range of public datasets.

4.4 End-to-End Speech Emotion Recognition Using Deep Neural Networks(Tzirakis et al., 2018)

The paper uses the continuous approach to SER. In the continuous approach, emotions are detected using 2 dimensions: arousal and valence. In the proposed model, 2 layers of LSTM were used on top of a CNN layer. The CNN layer used max pooling to reduce dimensionality. The researchers showed a relation between the kernel size and the pooling size which they called the rate of overlap. It was used to determine the kernel size of the convolution layers. CNN layers also had dropout for regularisation. The loss function used was based on concordance correlation coefficient. The paper tests on the RECOLA database which has 46 recordings of 300s and 16kHz. In training and testing, 20s clips were taken. The performance comparison increased from 0.753 (previous SOTA) to 0.787 for the arousal dimension and 0.43 to 0.44 for the valence dimension. The future scope of the paper was to try a deeper CNN and implement the model on a bigger dataset.

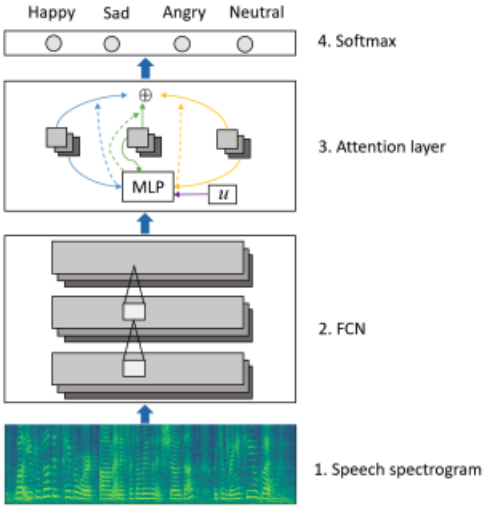
4.5 Attention Based Fully Convolutional Network for Speech Emotion Recognition(Zhang et al., 2018)

Fully Convolutional Neural Networks have been used to perform speech emotion recognition as they have found great applications in handling speech of variable length. In the proposed architecture, There is no need to perform segmentation on the speech data and this helps preserve information. Since emotions are evident only in specific parts of the speech data during long utterances, the attention mechanism in the proposed framework makes the

model aware of the relevant time-frequency region of the speech spectrogram.

4.5.1 Framework

The neural network takes the speech spectrogram as the input (without performing any padding or segmentation as mentioned in the abstract). The proposed neural network is also capable of handling spectrogram with variable sizes. The speech spectrogram is first encoded using a Fully Convolutional Network (the structure of which is described later), followed by an Attention layer which is capable of handling the variable relevant lengths of the time-frequency units, which is passed on to a softmax layer for classification (into the categories: happy, sad, angry, neutral) **FCN Encoder**



- Input: Speech Spectrogram (say of shape f x t x c; where f: frequency of spectrogram, t: time domains of the spectrogram, c: channel size of the spectrogram)
- Conv with kernel size 11, stride 4, number of channels = 96
- MaxPool (All maxpools having kernel size = 3, stride size = 2)
- Conv with kernel size 5 stride 1, number of channels = 256
- MaxPool
- Conv with kernel size 3, stride 1, number of channels = 384
- Conv with kernel size 3, stride 1, number of channels = 384
- Conv with kernel size 3, stride 1, number of channels = 256
- MaxPool
- Output: having size F x Tx C where F: frequency of spectrogram, T: time domains of the spectrogram, C: channel size of the spectrogram.

All convolutional layers are followed by ReLU activation function. The first two convolutional layers also have a local response normalisation. **Attention Layer**

- lambda is a hyperparameter which controls the uniformity of the importance weights of the annotation vectors. In this study, lambda was taken as 0.3 to achieve balance
- The scaled softmax function helps realise the idea that not all element vectors ei are contributing to the emotion state A in an equal manner.

4.6 Attending to Emotional Narratives(Wu et al., 2019)

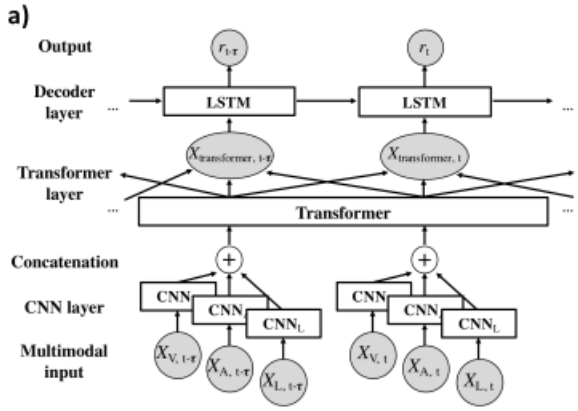
Over the course of the recent years, many frameworks have been proposed to work towards effective conversational agents that help engage in naturalistic emotion recognition. The current frameworks revolve around handling time-series data that allows for continuous emotion prediction over a period of time.
Often, when people utter a phrase or a sentence , especially during a long utterance, they tend to emphasise more on certain parts of the sentence by modulating the duration, speech, volume etc. and this too helps to convey the emotion of the speaker. The Attention mechanism helps to capture this intuition where the model tries to learn the relative importance of time-frequency units. Researchers have tried to capture this in frameworks involving applied RNNs, LSTM Models and employing Attention mechanisms in their deep learning frameworks to predict from complex time series data.
This paper explores attention mechanisms applied to perform emotion recognition on two narrative videos.
Two frameworks are suggested here which are based on Transformers, and the Memory Fusion Network suggested by Zadeh. et al. which capture the two key ideas of self-attention and cross-modality attention respectively. **Dataset used: The Stanford Emotional Narratives Dataset(SEND) are video recordings of participants narrating certain events.** Input:

Three modalities were chosen to be used from the video datasets-

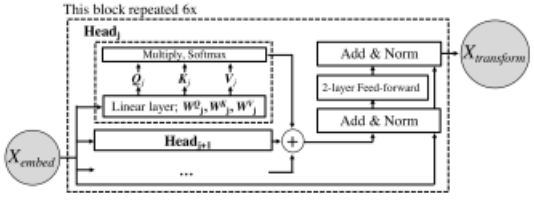
- Visual: Frames were taken every 0.1 second and 1000 features were extracted per frame from the final fully connected linear layer.
- Acoustic: 88 features were extracted for every second using openSMILE.
- Linguistic: Timestamps were assigned for each words to align the transcripts with the videos. Then, 300-dimensional GloVe word embeddings were used.

4.6.1 Simple Fusion Transformer(SFT: Self-Attention)

In the CNN Layer, The CNNs produce window-level embeddings for its corresponding modality for all time windows of a certain length. An input matrix is created by stacking the raw feature vectors from the the time-windows for each modality.

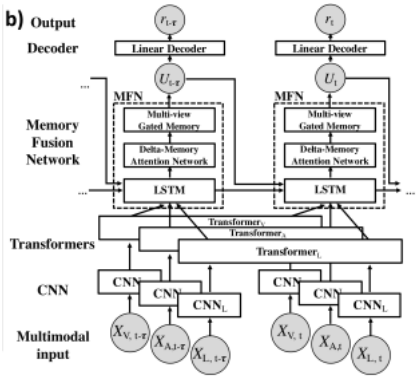


One dimensional CNN for each modality for each time-window with kernel-size = 2, followed by max pooling across the time dimension of the matrix.
This is followed by the transformer which serves as the encoder for the neural network. Multiple heads for attention are used in the transformer. The Architecture of the Transformer Used:



Each block consists of one multi-head attention layer and one feed-forward network.

4.6.2 Memory Fusion Transformer (MFT: Combining self attention with cross-modality attention)



Unlike The SFT Model which learns attention weights on time windows across all time but not across the different modalities within each time-window.
In the MFT, separate Transformer encoders are trained for each modality. The objective of the Memory Fusion Network is to learn the attention weights on the LSTM cells across two different time-windows.
DMAN: leans to attend to certain parts of the cell states in the LSTM.
MGM: propagates multimodal states over the subsequent time-windows.
In this way, the MFN that comprises of DMAN and MGM is able to apply attention across modalities as well as memory over the subsequent time states.

5 Framework

5.1 Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition

In this section of our two-fold study, we implemented the paper titled 'Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition' and further experimented with different pre-trained deep neural network architectures. In the paper, the authors have proposed an architecture using deep convolutional neural network, Bi-directional LSTM with Attention. The experiments in the paper were performed on EMO-DB and IEMOCAP database.

5.1.1 Pre-Processing

The dataset is standardised to have zero mean and unit variance. The data is augmented by speeding up the signals by a rate of 0.8 to 1.2. The signals are then split into frames using hamming window of 25 ms and 10 ms shift. Three channel of log Mel

Spectrograms are extracted using 64 Mel -Filter. This is done so that a three dimensional feature representation can be obtained by using the delta and the delta-deltas obtained. This is then sent to the pre-trained DCNN as the input.

5.1.2 Architecture

The DCNN model is pre-trained on ImageNet and used here to generate the segment-level features. The BiLSTM with Attention has been used for learning the higher-level features for temporal summarization which are relevant for us. The learned high-leve emotional features are then used in another deep neural network to perform the classification. The pre-trained DCNN architecture used in the paper was an AlexNet that had been trained for ImageNet.

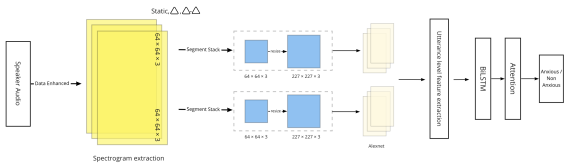


Figure 1: Architecture of first paper

5.2 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition

We implemented another paper titled "3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition".

5.3 Pre-Processing

Preprocessing has been done in a similar way as described in above paper, where log Mel with delta and delta-delta are used a model input. Using hamming window of 25 ms and 10 ms shift the audio is split into smaller signals. Log Mel is produced and its corresponding delta, delta-delta values. After computing the log-Mels with deltas and delta-deltas, we have obtained a 3-D feature representation as the CNN input.

5.4 Architecture

Convolutional Neural Network is combined with attention model to analyse 3-D log Mels for SER. CRNN (convolutional recurrent neural network) is used to extract high-level features for SER, these 3-D CNN sequential features are fed into LSTM for temporal summarization. Then, the attention layer takes a sequence of high-level features as input to produce utterance-level features. Finally, utterance-level features are used as the fully connected layer input to obtain higher-level features for SER.

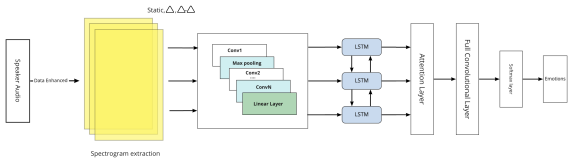


Figure 2: Architecture of second paper

6 Experiments

6.1 Paper I: Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition

We experimented with AlexNet as well as ResNet-18 architecture as the pre-trained DCNN in the model. Two datasets were used for each of the experiments as well, namely RAVDESS as well as Emo-DB. The authors of the paper had trained their model on EMO-DB and IEMOCAP. In the experiments that used AlexNet as the pre-trained dcnn model, SGD Optimiser was used to train the model. We achieved an accuracy within the range of 5% compared to the results from the paper with the pre-trained AlexNet architecture. Our results were therefore comparable with the results provided in the paper.

Table 3: Comparison of accuracies

Pretrain	EmoDB	RAVDESS
ResNet18	68	42
AlexNet	82.7	48
AlexNet(paper)	87.86	-

6.2 Paper 2 : 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition

Emotional Dyadic Motion Capture database IEMOCAP dataset was used to test the accuracy of the proposed model. IEMOCAP consists of 5 sessions and each session is displayed by a pair of speakers (female and male) in scripted and improvised scenarios. Both speakers have 10 speakers so 8 speakers were used for training purpose and remaining as test data. The test class distribution was imbalanced, so unweighted average recall (UAR) is used.

7 Results and analysis

7.1 Paper I: Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition

The results of the experiments performed on the EmoDB and RAVDESS dataset for both the architectures are as follows:

7.1.1 Results for experiments with AlexNet-based architecture

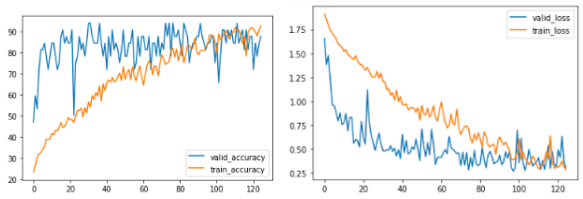


Figure 3: Accuracy and Loss Plots for AlexNet on the Emo DB Dataset

Table 4: Comparison of accuracies

Pretrain	EmoDB	RAVDESS
AlexNet	82.7	48
AlexNet(paper)	87.86	-

The accuracies of the models can be further improved by preprocessing the data to take its skewed nature into account.For example, In the EmoDB dataset, we can see that the samples for the emotion sad are far less than the rest of the emotions. In the RAVDESS dataset, we observed that the number of samples for disgust were considerably less than the number of samples for the other emotions in the dataset.

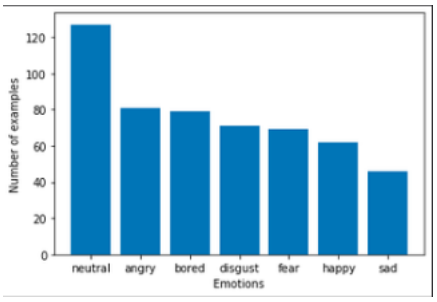


Figure 4: Exploratory Data Analysis on EmoDB dataset

7.1.2 Results for experiments with ResNet-18 based architecture

The low accuracy scores for the RAVDESS dataset can be improved by using a bigger ResNet like the ResNet-50 or the ResNet-101.

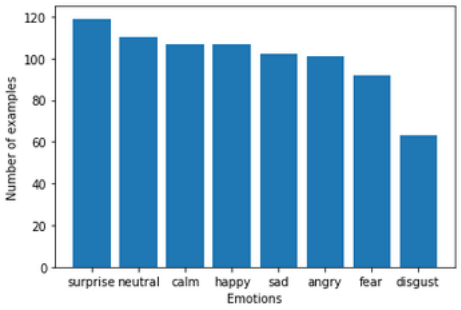


Figure 5: Exploratory Data Analysis on RAVDESS dataset

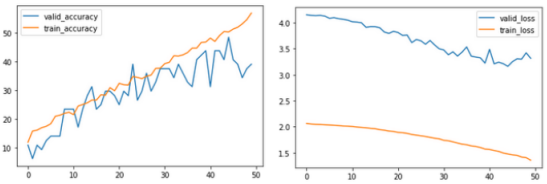


Figure 6: Accuracy and Loss Plots for ResNet-18 on the RAVDESS dataset

Table 5: Comparison of accuracies

Pretrain	EmoDB	RAVDESS
ResNet-18	68	42

7.2 Paper 2 : 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition

The paper reached highest recall score of 64.74±5.44 on IEMOCAP dataset. They used 6 convolutional layers and LSTM layer. We were able to produce similar results of 65 recall score. We used 7 convolutional and LSTM layer to reproduce the result. Sad has obtained the highest recognition rate while happy emotion has obtained the lowest recognition rate.

8 Individual Contribution

Equal contribution from each member. We have implemented 2 different papers. The detail contribution is as follows.

- Joykirat : Literature Review, Preprocessing of paper 1 and model of paper 2.
- Sudarshan : Literature Review, Preprocessing and model of paper 1.

	Precision	Recall	f1-score	support
0	0.53	1.00	0.69	18
1	0.64	0.78	0.70	73
2	0.31	0.18	0.23	56
3	0.69	0.66	0.67	151
accuracy			0.62	298
macro avg	0.54	0.65	0.57	298
weighted avg	0.60	0.62	0.60	298

Figure 7: Paper 2 Classification report

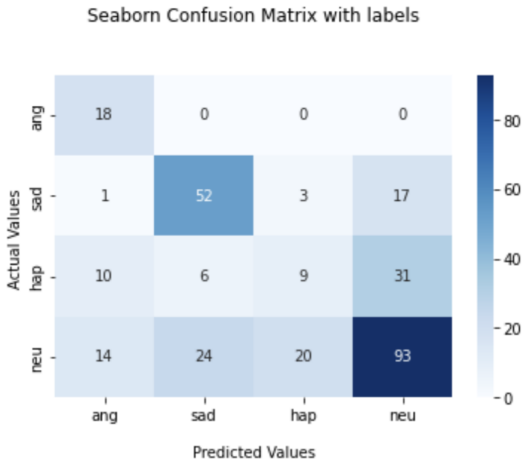


Figure 8: Paper 2 Confusion matrix

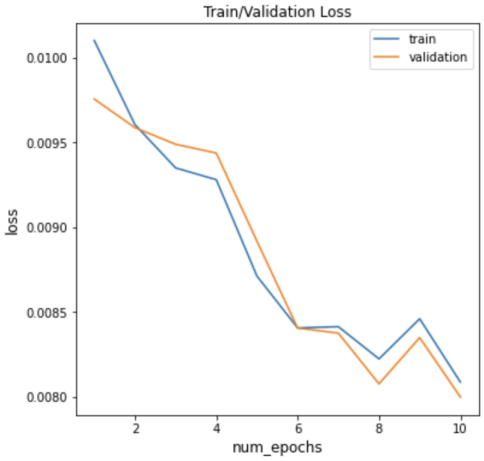


Figure 9: Paper 2 train validation loss

- Rohan : Literature Review, Preprocessing of paper 2 and model for paper 1.

References

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Suraj Tripathi, Abhay Kumar, Abhiram Ramesh, Chirag Singh, and Promod Yenigalla. 2019. Deep learning based emotion recognition system using speech features and transcriptions. *arXiv preprint arXiv:1906.05681*.

Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. 2018. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5089–5093. IEEE.

Xianfeng Wang, Min Wang, Wenbo Qi, Wanqi Su, Xiangqian Wang, and Huan Zhou. 2021. A novel end-to-end speech emotion recognition network with stacked transformer layers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6289–6293. IEEE.

Zhengxuan Wu, Xiyu Zhang, Tan Zhi-Xuan, Jamil Zaki, and Desmond C Ong. 2019. Attending to emotional narratives. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 648–654. IEEE.

Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu. 2018. Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1771–1775. IEEE.