Supplementary File

Towards the explainability of Multimodal Speech Emotion Recognition

Puneet Kumar[†]*, *Vishesh Kaushik*[‡]*, *Balasubramanian Raman*[†]

[†]Computer Science and Engg. Dept., Indian Institute of Technology, Roorkee, India, 247667
[‡]Mechanical Engg. Dept., Indian Institute of Technology, Kanpur, India, 208016
pkumar99@cs.iitr.ac.in, kvishesh@iitk.ac.in, bala@cs.iitr.ac.in

1. Architecture of the Baseline Model

The baseline method's architecture is described in Fig. 1. It encodes speech and text information separately, and then their concatenated vector is used to train a neural-network-based classifier.

1.1. Proposed Methodology

1.1.1. Speech Emotion Recognition Phase

This phase takes input as the mean values of Mel-Spectrogram, Mel-frequency cepstral coefficients (MFCCs), and Chroma features. A 128-dimensional Mel-spectrogram, a 40-dimensional MFCC, and a 12-dimensional chroma vector were concatenated to construct a 180-dimensional input feature vector. The feature vector was passed into a dense neural network followed by a dropout layer. This phase returned a 470-dimensional speech encoding vector S.

1.1.2. Text Emotion Recognition Phase

In this phase, corresponding text transcription was provided as input. We used 50-dimensional Global Vectors (GloVe) word embedding [1] for each token. The 20×50 dimensional text input was formed and passed to an LSTM network followed by a flatten layer and a dropout layer. A text encoding vector T of 1000 dimensions was obtained as the output.

1.1.3. Multimodal Emotion Recognition Phase

The speech and text encoding vectors are concatenated and passed to a softmax layer to get the predictions as per Eq. 1. We have used the sparse categorical cross-entropy loss function and Adam optimizer. The model is updated in response to the loss function's output in every iteration. Here i is the emotion class, y_i is the prediction for i^{th} class, W is the weight matrix, and b is the bias term.

$$y_i = softmax(concat(S,T)^T W + b) \quad (1)$$

2. More Ablation Studies

More ablation studies have been performed on the conversation sessions of the IEMOCAP dataset [2].

2.1. Deciding Baseline Model's Architecture

For the speech module, the Fully Connected (FC) network performed better than Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Two and three-layered FC networks (FC-2 and FC-3) performed better than the networks with more layers. The best performance was obtained with sessions 1 and 2 for FC-2 for the speech phase. RNNs with Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU) units were evaluated for the text module. As shown in Table 1, LSTM with two layers (LSTM-2) performed better than the other network choices.

Table 1: *Ablation study to select baseline model's architecture*

Speech Module	Text Module	Accuracy
FC-2	GRU-2	69.45%
FC-2	GRU-3	67.20%
FC-2	LSTM-2	**72.30**%
FC-2	LSTM-3	70.75%
FC-3	GRU-2	70.69%
FC-3	GRU-3	68.12%
FC-3	LSTM-2	70.85%
FC-3	LSTM-3	70.00%

2.2. Deciding Proposed Model's Architecture

The speech module of the proposed system uses GRU and attention on the speech features. Three speech features, i.e., mel-spectrogram, chroma, and mel-frequency cepstral coefficients (MFCC), are available. We have experimented by using the attention ('Att') with various combinations of these features. The best performing configuration for the speech module is chosen, and then the experiments to determine the text module's configuration have been performed.

Table 2: *Ablation study to select proposed model's architecture.*

Speech Module	Text Module	Accuracy
FC-2 + No attention	LSTM-2	72.30%
FC-2 + No attention	BERT	**73.40**%
GRU + Att. with chroma only	BERT	67.55%
GRU + Att. with mfcc only	BERT	68.08%
GRU + Att. with mel-spectrogram only	BERT	69.68%
GRU + Att. with chroma and mfcc	BERT	70.21%
GRU + Att. with chroma and mel-sp	BERT	71.81%
GRU + Att. with mfcc and mel-sp	BERT	73.40%
GRU + Att. with all three	BERT	**75.00**%
LSTM + Att. with all three	BERT	74.46%
RNN + Att. with all three	BERT	73.94%

For the text module, we have first experimented with LSTM-2 (referring to the baseline model's architecture) and Bidirectional Encoder Representations from Transformer

* Denotes Equal Contribution.

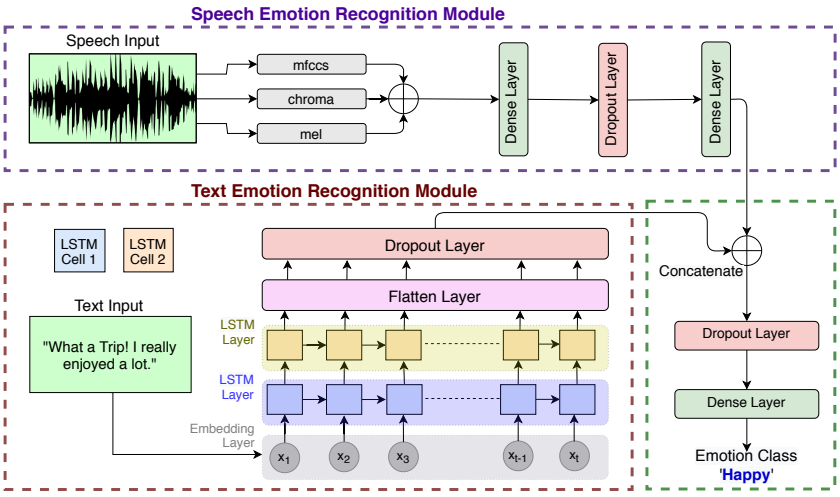


Figure 1: *Schematic description of the baseline method's architecture. Here, x_i is i^{th} token of the text input.*

(BERT). BERT has been chosen for the final implementation based on its performance. Then, GRU, Simple RNN and LSTM have been analyzed for the speech module, and GRU has been chosen as it has outperformed LSTM and Simple RNN. The summary of this study has been presented in Table 2.

2.3. 'Audio-only' vs 'Audio + Text' experiments

The proposed method and the baseline method are implemented with the architectures determined in Section 2.1 and Section 2.2 respectively. The emotion recognition accuracy has been observed using only speech modality and then using speech and text modality both. The same has been summarized in Table 3.

Table 3: *Ablation Study for audio-only vs A+T (Audio + Text)*

Method	Speech Module	Text Module	Accuracy
Baseline	FC-2	–	68.45%
	FC-2	LSTM-2	72.30%
Proposed	GRU + Attention	–	70.20%
	GRU + Attention	BERT	75.00%

The proposed model and the baseline model have performed better on using speech and text modalities compared to using only speech modality. It proves the applicability of using the information from text modality to improve emotion recognition in speech.

3. Experiments and Results on More Datasets

Results for speech-only datasets have been included here.

3.1. RAVDESS Dataset

The proposed system has also been evaluated with Ryerson Audio Visual Data of Emotional Speech and Song (RAVDESS) dataset [3] containing 7,356 speech utterances spoken by 24 speakers.

3.1.1. Accuracy & Confusion Matrix

Fig. 2 shows the confusion matrix for the proposed method. It has resulted in an unweighted emotion recognition accuracy of 90.16% and weighted accuracy of 83.24%.

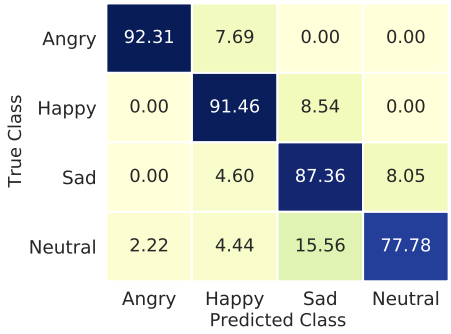


Figure 2: *Confusion matrix for RAVDESS dataset*

Table 4: *Result comparison for RAVDESS dataset*

Method	Author	Accuracy
Convolutional Neural Network	Issa et al. [4]	71.61%
Bagged Ensemble of SVMs	Bhavan et al. [5]	75.69%
Clustering based SER	Mustaqeem et al. [6]	77.02%
Multi Task Hierarichel SVM	Zhang et al. [7]	83.15%
Artificial Neural Network	Tomba et al. [8]	89.16%
Proposed Method		**90.16%**

3.1.2. Comparison with existing results

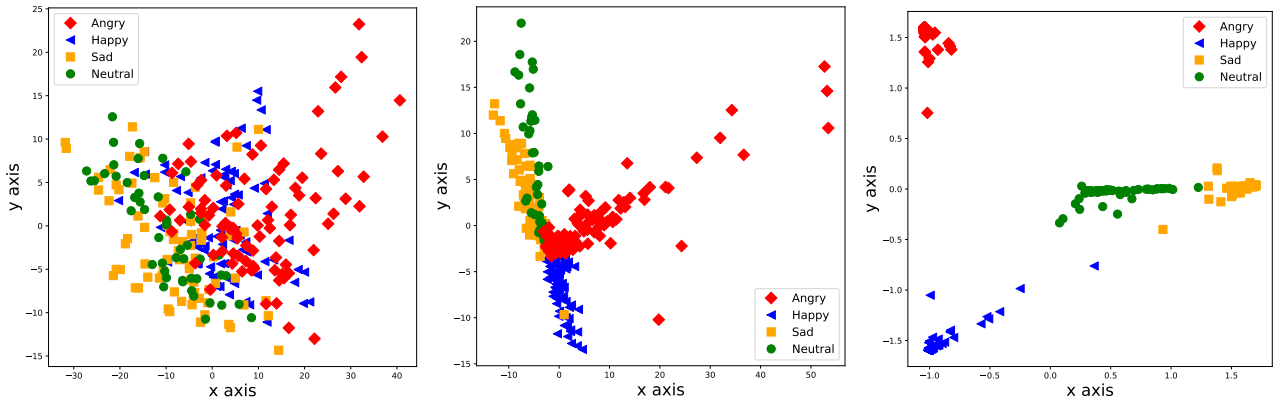
Table 4 compares the performance of the proposed model on RAVDESS dataset with the state-of-the-art (SOTA) SER approaches. The unweighted accuracy has been considered for the comparison.

3.1.3. Embedding Plots

The emotion embedding plots for the model trained on RAVDESS dataset have been visualized in Fig. 3.

3.1.4. Intersection Matrix

Table 5 shows the intersection matrices for the last three layers of the trained network.



(a) Third-last Layer (b) Second-last Layer (c) Output Layer

Figure 3: *Emotion embedding plots for RAVDESS dataset*

Table 5: *Intersection matrices for RAVDESS dataset*

	Third-last Layer				**Second-last Layer**				**Output Layer**			
	Angry	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*
Angry	1.000	0.678	0.505	0.561	1.000	0.514	0.378	0.461	1.000	0.000	0.000	0.000
Happy		1.000	0.357	0.377		1.000	0.284	0.246		1.000	0.000	0.000
Sad			1.000	0.542			1.000	0.341			1.000	0.000
Neutral				1.000				1.000				1.000

3.2. MSP-IMPROV Dataset

The experiments have also been performed for MSP-IMPROV dataset [9] containing 8,438 samples.

3.2.1. Accuracy & Confusion Matrix

Fig. 4 shows the confusion matrix for the proposed method. It has resulted in an unweighted emotion recognition accuracy of 62.95% and weighted accuracy of 57.32%.

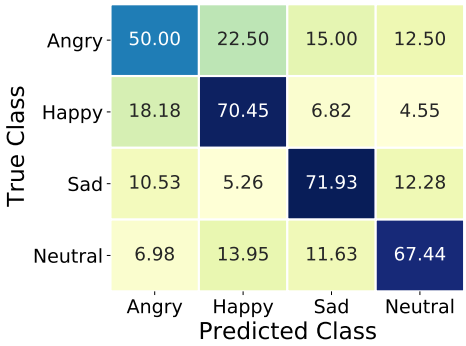


Figure 4: *Confusion matrix for MSP-Improv dataset*

3.2.2. Comparison with existing results

Table 6 compares the performance of the proposed model on MSP-IMPROV dataset with the state-of-the-art (SOTA) SER approaches. The unweighted accuracy has been considered for the comparison.

Table 6: *Result comparison for MSP-IMPROV dataset. Here 'ASRNN': Attention-based Sliding Recurrent Neural Networks, 'CRNN': Convolutional Recurrent Neural Network.*

Method	**Author**	**Accuracy**
Attentive CNN	Michael et al. [10]	45.76%
CRNN	Siddique et al. [11]	52.43%
CNN	Zakaria et al. [12]	53.8%
ASRNN	Peng et al. [13]	55.7%
ProgNet (Transfer Learning)	Gideon et al. [14]	60.5%
Proposed Method		**62.95%**

3.2.3. Embedding Plots

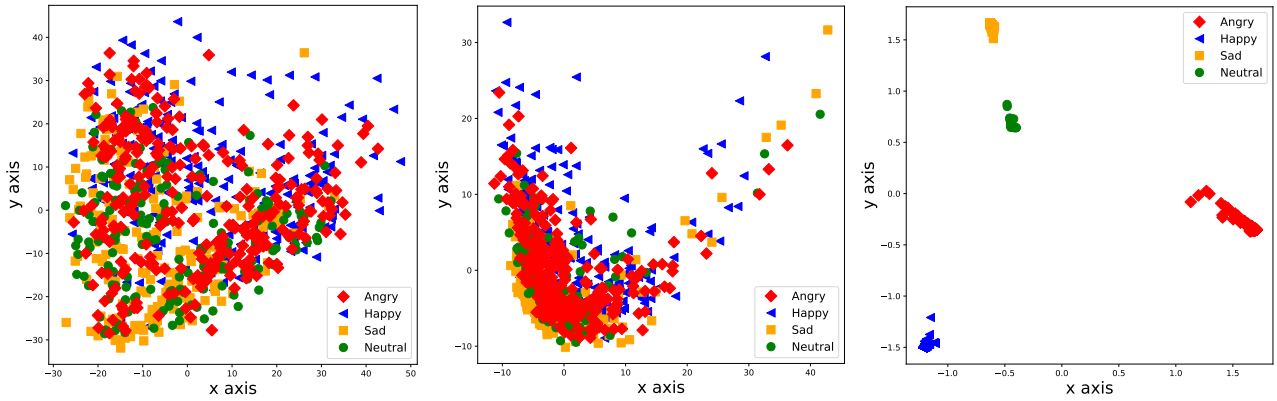
Fig. 5 shows the emotion embedding plots for the model trained on the MSP-IMPROV dataset.

3.2.4. Intersection Matrix

Table 7 shows the intersection matrices for the last three layers of the trained network.

Discussion: By monitoring Fig. 3 and Table 5 simultaneously, it can be observed that the convergence of the intersection values is in correspondence to the convergence of the embedding plots. Similar inference can be made by analysing Fig. 5 and Table 7 side by side. It has been observed that the intersection values converge faster for speech-only datasets as compared to the multimodal datasets.

The experiments to determine the number and size of the network layers for IEMOCAP dataset have been described in Section 3.1.2 of the main paper. This concept is applicable for RAVDESS and MSP-IMPROV datasets in a similar way.



(a) Third-last Layer (b) Second-last Layer (c) Output Layer

Figure 5: *Emotion embedding plots for MSP-IMPROV dataset*

Table 7: *Intersection matrices for MSP-IMPROV dataset*

	Third-last Layer				**Second-last Layer**				**Output Layer**			
	Angry	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*
Angry	1.000	0.018	0.011	0.060	1.000	0.015	0.003	0.027	1.000	0.000	0.000	0.000
Happy		1.000	0.008	0.052		1.000	0.001	0.024		1.000	0.000	0.000
Sad			1.000	0.10			1.000	0.009			1.000	0.000
Neutral				1.000				1.000				1.000

4. References

[1] J. Pennington, R. Socher, and C. D. Manning, "GLOVE: Global Vectors for word representation," in *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[2] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive Emotional dyadic MOtion CAPture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[3] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English," *PloS one*, vol. 13, no. 5, 2018.

[4] D. Issa, M. F. Demirci, and A. Yazici, "Speech Emotion Recognition with Deep Convolutional Neural Networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[5] A. Bhavan, P. Chauhan, R. R. Shah *et al.*, "Bagged Support Vector Machines for Emotion Recognition from Speech," *Knowledge-Based Systems*, vol. 184, p. 104886, 2019.

[6] M. Sajjad, S. Kwon *et al.*, "Clustering-based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," *IEEE Access*, vol. 8, pp. 79 861–79 875, 2020.

[7] B. Zhang, G. Essl, and E. M. Provost, "Recognizing Emotion from Singing and Speaking using Shared Models," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 139–145.

[8] K. Tomba, J. Dumoulin, E. Mugellini, O. A. Khaled, and S. Hawila, "Stress Detection Through Speech Analysis." in *International Conference on Emerging Trends in Engineering (ICETE)*, 2018, pp. 560–564.

[9] C. Busso et al., "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Transactions on Affective Computing (TAC)*, vol. 8, no. 1, pp. 67–80, 2016.

[10] M. Neumann and N. T. Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7390–7394.

[11] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct Modelling of Speech Emotion from Raw Speech," *arXiv preprint arXiv:1904.03833*, 2019.

[12] Z. Aldeneh and E. M. Provost, "Using Regional Saliency for Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2741–2745.

[13] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech Emotion Recognition using 3D Convolutions and Attention-based Sliding Recurrent Networks with Auditory Front-ends," *IEEE Access*, vol. 8, pp. 16 560–16 572, 2020.

[14] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive Neural Networks for Transfer Learning in Emotion Recognition," *arXiv preprint arXiv:1706.03256*, 2017.