INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

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

Abstract

In this paper, a multimodal speech emotion recognition system has been developed, and a novel technique to explain its predictions has been proposed. The audio and textual features are extracted separately using attention-based Gated Recurrent Unit (GRU) and pre-trained Bidirectional Encoder Representations from Transformers (BERT), respectively. Then they are concatenated and used to predict the final emotion class. The weighted and unweighted emotion recognition accuracy of 71.7% and 75.0% has been achieved on Emotional Dyadic Motion Capture (IEMOCAP) dataset containing speech utterances and corresponding text transcripts. The training and predictions of network layers have been analyzed qualitatively through emotion embedding plots and quantitatively by analyzing the intersection matrices for various emotion classes' embeddings.
Index Terms: Multimodal emotion recognition, deep network explainability, intersection matrix, embedding plot.

1. Introduction

The need to develop multimodal speech processing systems capable of recognizing various emotions from speech is rapidly increasing [1]. Human emotions and intentions are well contained in the speech. Speech Emotion Recognition (SER) has a wide range of applications such as robotics, security, language translation, automated identification, intelligent toys, and lie detection [2]. Despite many developments in speech processing, natural emotion understanding is still a challenging task for computational systems [3].

Several attempts have been made to classify the emotional content portrayed by speech utterances [4, 5, 6]. In the context of using state-of-the-art Deep Neural Networks (DNNs) for SER, Majumder et al. [7] implemented an attention-based approach to track speakers' identities showing specific emotions. Joint Feature Learning-based approaches have also been used for SER [8, 9, 10]. Researchers have attempted to use the information from text modality to improve SER [11, 12]. For example, Siriwardhana et al. [13] used transformer-based pre-trained models to fuse the information from text and speech modalities. In another work, H. Feng [8] combined an automatic speech recognition module along with the SER module. However, fine-tuning the pre-trained models for SER is a complex process, and it needs to be explored further [14].

Attention-based deep-learning approaches have started finding their application for SER [15, 16, 17]. In this regard, M. Xu [18] used multi-head self-attention and observed an increased performance. In another work, Seyedmahdad et al. [19] implemented local-attention to learn the emotion features automatically. However, the aforementioned deep-learning-based systems act as black-box. It is not possible to understand their internal mechanism [20].

[*] equal contribution

Explaining the internal mechanism of deep-learning-based classification methods has emerged as a recent research topic [21]. In this context, Lin et al. [22] proposed a method for interpreting multimodal emotion recognition of biological signals using deep-learning. In another work, Zhang et al. [23] analyzed the effect of various audio features on emotion arousal and interpreted the corresponding response. Riberio et al. [24] developed a technique to find the input's part responsible for a particular output. In another work, Shrikumar et al. [25] came up with a method to break down the output predictions by tracing the contributions of all the neurons. However, the above-discussed methods were unable to show the network's layer-by-layer training. It inspired us to develop a method to explain and interpret a DNN based SER system's predictions.

The proposed system encompasses a speech emotion recognition (SER) module and a text emotion recognition (TER) module. For the SER module, Gated Recurrent Units (GRUs) [26] have been implemented along with attention to extract the audio features. For the TER module, a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model [27] has been used for textual feature extraction. The audio and textual features have been concatenated and used for the prediction of the final emotion class. A novel approach to explain the mechanism of the emotion recognition model has also been proposed. The proposed approach has achieved weighted and unweighted accuracy of 71.7% and 75.0%. The proposed system's training and predictions have been explained by observing the intersection matrix's values for the intermediate layers and the output layer. The code and supplementary file are available $github.com/MIntelligence\text{-}Group/SpeechText\_EmoRec$.

The contributions of the paper are as follows. A deep-learning-based multimodal speech emotion recognition system has been developed that uses the corresponding text along with speech utterances. Then, a novel technique to explain the proposed system's predictions has been proposed. We have defined the *Intersection Matrix* to calculate the intersection among the embeddings of various emotion classes. The emotion recognition mechanism has been qualitatively explained through emotion embedding plots. The proposed network's training has been quantitatively interpreted by analyzing the intersection of various emotion classes in terms of the Intersection Matrix.

2. Multimodal Emotion Recognition

2.1. Proposed Methodology

The proposed system's architecture is described in Fig. 1 and explained in the following sections.

2.1.1. Speech Emotion Recognition Phase

This phase takes a 128-dimensional mel-spectrogram (m), a 40 dimensional mel-frequency cepstral coefficients (MFCC, f),

http://dx.doi.org/10.21437/Interspeech.2021-1718

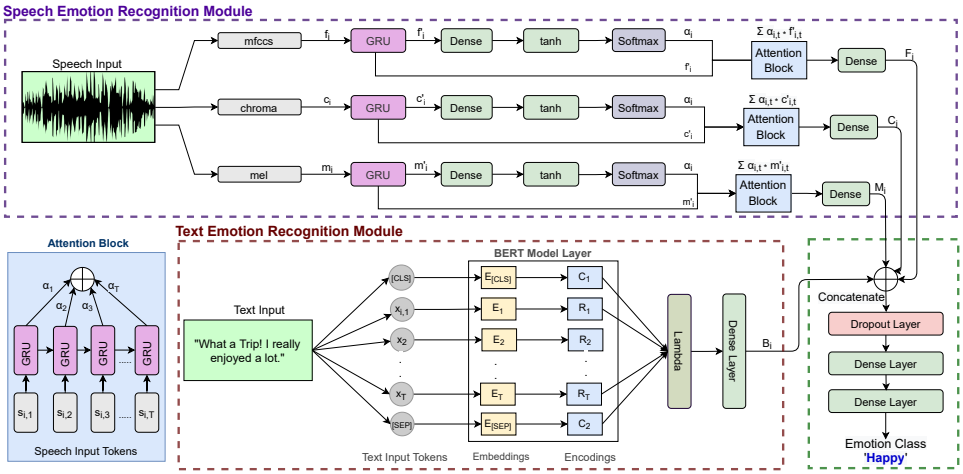


Figure 1: *A description of the proposed method's architecture. Here, x_i is i^{th} token of text input and s_i is i^{th} token of speech input.*

and a 12-dimensional chroma (c) vectors as input. The attention mechanism has been implemented separately for each of these features to extract their parts with the most significant emotional information. For i^{th} input sample, the speech feature vector s_i is passed through the GRU layer for T time-steps to obtain s'_i, where t denotes t^{th} time-step.

$$\{s'_{i,t}\}_{t=1}^{t=T} = GRU(s_i) \qquad \forall s \in \{m, f, c\} \tag{1}$$

The hidden representation $h_{i,t}$ of the speech tokens $s_{i,t}$ is computed in Eq.2. Then the similarity of $h_{i,t}$ is measured with context vector s_w as the importance of t^{th} speech part. The normalized importance weight $\alpha_{i,t}$ is obtained in Eq.3. Here, ^{T} denotes transpose.

$$h_{i,t} = tanh(W_s s'_{i,t} + b_s) \tag{2}$$

$$\alpha_{i,t} = \frac{exp(h_{i,t}^T s_w)}{\sum_t exp(h_{i,t}^T s_w)} \tag{3}$$

Finally, the speech vector a_i is calculated as the weighted sum of all the speech tokens based on the weights $\alpha_{i,t}$.

$$a_i = \sum_t \alpha_{i,t} s'_{i,t} \tag{4}$$

Speech vector a_i is passed through a dense layer to get the speech vectors C_i, M_i and F_i for Chroma and Mel-Spectrogram and MFCC features respectively.

2.1.2. Text Emotion Recognition Phase

In this phase, the text transcriptions corresponding to the speech utterances fed to SER phase are provided as the input to a pre-trained Bidirectional Encoder Representations from Transformer (BERT) model with 12 transformer blocks, 12 attention heads and 768 hidden units [27]. For the i^{th} input token, it produces the encoded hidden vectors R_i for the input embeddings E_i. The encodings for the special tokens [CLS: Classification] and [SEP: Separator] are denoted as C_1 and C_2. $\{x_{i,1}, x_{i,2} \dots x_{i,T}\}$ are the input tokens for i^{th} data point for T time-steps while B_i denotes the text feature vector extracted by BERT module for i^{th} input token.

2.1.3. Multimodal Emotion Recognition Phase

Then we have concatenated all the four vectors C_i, M_i, F_i, B_i and passed the concatenated layer through a dense layer which gives us E_i. Then the final vector E_i is passed to a softmax layer to get the predictions as Eq given below. We have used the *sparse categorical cross-entropy* as the loss function and *Adam* as the optimizer. The model is updated in response to the loss function's output in every iteration. Here e is the emotion class, y_e is the prediction for e^{th} class, W is the weight matrix, and b is the bias term.

$$y_e = softmax(W^T E_i + b) \tag{5}$$

2.2. Explainabilty of Multimodal Emotion Recognition

The proposed approach explains the mechanism of the emotion recognition model by observing emotion clusters for various layers. It contains the following steps:

i) For every emotion class i, compute the principle components **(x_i, y_i, z_i)** of the embedding of p^{th} layer from the end.

ii) For every emotion class i and (x_i, y_i, z_i) components, mean m_i and standard deviation $\sigma_{\bar{m}_i}$ are calculated as follows, where n_i is the no. of samples for i^{th} class and k is the k^{th} data point.

$$\bar{m}_i = \frac{1}{n_i} \sum_{k=1}^{k=n_i} m_k; \;\; \sigma_{\bar{m}_i} = \frac{1}{n_i} \sum_{k=1}^{k=n_i} (m_k - \bar{m}_i)^2 \tag{6}$$

iii) $[L_i(m), R_i(m)]$ is defined as the range for i^{th} emotion where $R_i(m)$ and $L_i(m)$ are the right and left extreme points for i^{th} emotion class's m^{th} component's spread. The extreme points are defined using Eq. 7 and range is defined for every principle component.

$$L_i(m) = \bar{m}_i - 2\sigma_{\bar{m}_i} \;\; R_i(m) = \bar{m}_i + 2\sigma_{\bar{m}_i} \;\; \forall m \in \{x, y, z\} \tag{7}$$

iv) Eq. 8 computes the intersection between emotion classes i and j, i.e., $I_{i,j}(m)$. Here, $I_{i,j}(m)$ is the intersection between spread of m^{th} component's data for emotion classes i and j.

$$I_{i,j}(m) = \frac{max\{(min(R_i(m), R_j(m)) - max(L_i(m), L_j(m))), 0\}}{max(R_i(m), R_j(m)) - min(L_i(m), L_j(m))} \tag{8}$$
$$\forall m \in \{x, y, z\}, \quad \forall i, j \in \{1, 2, 3, 4\}$$

v) The total intersection between two emotion classes i and j is denoted as $I_{i,j}$. As shown in Eq. 9, it is calculated as the product of all component-wise intersections between emotion classes i and j. The values for $I_{i,j}$ are in the range [0,1]. It has a maximum value of one when $i = j$, i.e., when the spread of m^{th} component's data is same for two emotion classes.

$$I_{i,j} = I_{i,j}(x) * I_{i,j}(y) * I_{i,j}(z) \tag{9}$$

vi) Finally, the *Intersection Matrix* of size 4×4 is defined as I. The $(i, j)^{th}$ element of I represents the total intersection $I_{i,j}$

between the emotion classes i and j. Lower value of $I_{i,j}$ for $i \neq j$ corresponds to better classification of the elements of i^{th} and j^{th} emotion classes into different emotion clusters. Fig. 2 shows an example of intersection calculation steps to demonstrate the aforementioned concept.

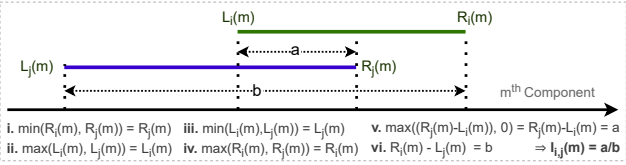


Figure 2: *Example calculation of $I_{i,j}(m)$ for m^{th} component in emotion class i, j. Here $L_k(m)$ and $R_k(m)$ denote left and right extreme of k^{th} emotion's m^{th} component's data spread.*

3. Experiments and Results

3.1. Implementation

3.1.1. Dataset and Training Strategy

We have trained and evaluated the proposed system on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [28] containing speech utterances and corresponding text transcriptions. Four emotion classes, i.e., 'angry,' 'happy,' 'sad,' and 'neutral,' have been used to compare with the previous works that used the IEMOCAP dataset. The speech samples marked as 'excited' have been merged with the samples marked as 'happy.' The merging of emotion labels has been done based on Plutchik's wheel of emotions [29]. The training-testing split of 80%-20% and 5-fold cross-validation has been used.

3.1.2. Ablation Study

The ablation study has been performed to select the appropriate architecture for the proposed model. The intersection matrix is used to determine the number and size of its dense layers.

a) Deciding the number of layers

Table 1 and 2 depict the intersection matrices of the models when two and three dense layers are used along with a concatenation layer. Both the models are trained for the same number of epochs, and the convergence of intersection matrices' values is observed. As observed from Table 2, adding one more layer has degraded the overall performance as the values in intersection matrices are larger than the values in Table 1. It suggests using two dense layers along with a concatenation layer.

b) Deciding the size of the layers

We have experimented with different sizes for the second-last layer by keeping the dimensions of other layers fixed. We have considered three different sizes, i.e., 400, 600, and 800-dimensional layers. Table 1, 3 and 4 show the intersection matrices for last three layers for the models with 600, 400 and 800 dimensional second-last layer. The smaller intersection values for the last layer in Table 1 suggest using a 600-dimensional second-last layer for faster convergence. It has been implemented for the final model containing two dense layers after the concatenation layer.

3.2. Results & Evaluation

The experimental results for IEMOCAP dataset are discussed as follows, while the results for two more datasets (RAVDESS [30] and MSP-IMPROV [31]) are included in the supplementary file.

3.2.1. Accuracy & Confusion Matrix

Fig. 3 shows the confusion matrix for the proposed method. It has resulted in a unweighted emotion recognition accuracy of 75% and weighted accuracy of 71.7%.

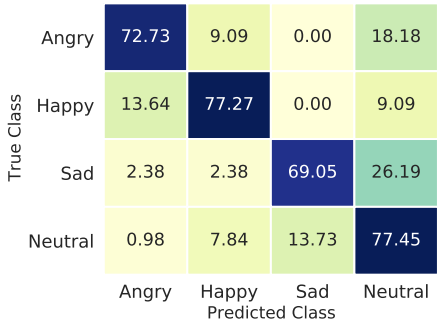


Figure 3: *Confusion matrix*

3.2.2. Embedding Plots & Intersection Matrix

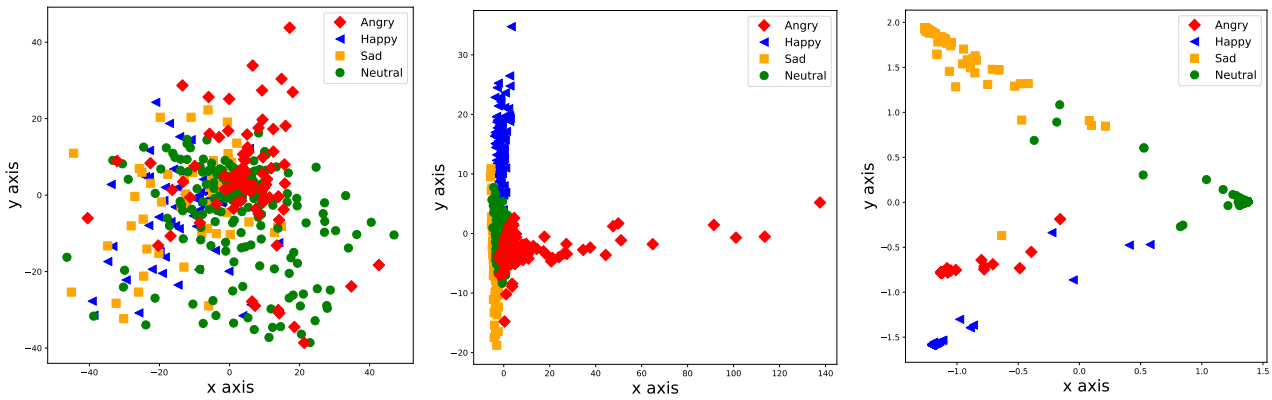
The emotion embedding plots for the trained model have been visualized in Fig. 4. The Intersection matrix has been used to explain the emotion classification mechanism of the proposed method. Table 5 shows the intersection matrices for the last three layers of the trained network. The clusters of the same emotions should overlap, and those for different emotions should be separated. The value of one for the diagonal elements and convergence of non-diagonal elements' values to zero validates this hypothesis. Inter-emotion clusters' separability improves from intermediate to output layers. The decrease in non-diagonal elements' values in the intersection matrix affirms that the model has learned to classify various emotions.

3.2.3. Comparison with State-of-the-art results

Table 6 compares the performance of the proposed model with the baseline method and state-of-the-art (SOTA) SER approaches. The baseline is an end-to-end multimodal framework containing dense layers without attention for the SER phase and Long Short Term Memory (LSTM) based DNN for the TER phase. The complete architecture of the baseline model and the ablation study to determine it has been included in the supplementary file. For SOTA, various methods that have reported the best emotion recognition accuracy for IEMOCAP dataset have been considered.

Table 6: *Result comparison with state-of-the-art methods. 'UA' and 'WA' denote the Unweighted and the Weighted Accuracies; 'A' and 'T' represent Speech and Text modalities respectively.*

Method	**Modality**	**UA**	**WA**
Pre-trained Transformer [17]	A	—	61.80%
RNN + Attention [7]	A	—	62.75%
RNN + Local Attention [19]	A	63.50%	58.80%
Memory Network [6]	A+T	63.80%	—
Self-attention for SER [15]	A	65.40%	66.90%
Attention Head Fusion [18]	A	67.28%	67.94%
Feature Learning [10]	A	68.10%	63.80%
End-to-end SER [8]	A+T	69.70%	68.60%
Transfer Learning [5]	A+T	68.70%	—
LSTM + GloVe [12]	A+T	69.74%	—
Baseline Method (ours)	A+T	72.3%	63.25%
Proposed Method (ours)	A+T	**75.00%**	**71.70%**



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

Figure 4: *Emotion embedding plots for various layers*

Table 1: *Intersection matrices on using three dense layers with 600-dimensional second-last layer*

	Third-last Layer				**Second-last Layer**				**Output Layer**			
	Angry	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*
Angry	1.0000	0.0565	0.0015	0.0201	1.0000	0.0063	0.0004	0.0076	1.0000	.0045	0.0001	0.0042
Happy		1.0000	0.0029	0.1248		1.0000	0.0035	0.0249		1.0000	0.0010	0.0143
Sad			1.0000	0.0387			1.0000	0.0255			1.0000	0.0030
Neutral				1.0000				1.0000				1.0000

Table 2: *Intersection matrices on using four dense layers with 600-dimensional second-last layer*

	Fourth-last Layer				**Third-last Layer**				**Second-last Layer**				**Output Layer**			
	Angry	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*
Angry	1.0000	0.0276	0.0059	0.0300	1.0000	0.1189	0.0017	0.0054	1.0000	0.0417	0.0130	0.0705	1.0000	0.0383	0.0000	0.0278
Happy		1.0000	0.1609	0.0401		1.0000	0.0034	0.0131		1.0000	0.0127	0.0732		1.0000	0.0000	0.1024
Sad			1.0000	0.0292			1.0000	0.0292			1.0000	0.0221			1.0000	0.022
Neutral				1.0000				1.0000				1.0000				1.0000

Table 3: *Intersection matrices on using three dense layers with 400-dimensional second-last layer*

	Third-last Layer				**Second-last Layer**				**Output Layer**			
	Angry	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*
Angry	1.0000	0.1893	0.0268	0.02493	1.0000	0.1117	0.0036	0.2926	1.0000	0.1969	0.0000	0.0681
Happy		1.0000	0.0078	0.1592		1.0000	0.0004	0.0618		1.0000	0.0211	0.3021
Sad			1.0000	0.0409			1.0000	0.0042			1.0000	0.0698
Neutral				1.0000				1.0000				1.0000

Table 4: *Intersection matrices on using three dense layers with 800-dimensional second-last layer*

	Third-last Layer				**Second-last Layer**				**Output Layer**			
	Angry	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*
Angry	1.0000	0.3534	0.0014	0.0101	1.0000	0.3534	0.0014	0.0101	1.0000	0.1666	0.0396	0.1187
Happy		1.0000	0.0011	0.0141		1.0000	0.0011	0.0141		1.0000	0.0554	0.3847
Sad			1.0000	0.0832			1.0000	0.0832			1.0000	0.0971
Neutral				1.0000				1.0000				1.0000

Table 5: *Intersection matrices for the final proposed model*

	Third-last Layer				**Second-last Layer**				**Output Layer**			
	Angry	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*
Angry	1.0000	0.0023	0.0014	0.0042	1.0000	0.0022	0.0047	0.0049	1.0000	0.0000	0.0000	0.0000
Happy		1.0000	0.0399	0.0453		1.0000	0.0242	0.0458		1.0000	0.0000	0.0000
Sad			1.0000	0.0280			1.0000	0.0179			1.0000	0.0001
Neutral				1.0000				1.0000				1.0000

4. Conclusion and Future Work

A deep-learning-based system has been developed to recognize emotions from speech utterances and corresponding text. It has performed better than the existing SER methods. A novel technique has been proposed to explain the training and predictions of the proposed system. It helped in understanding the convergence of emotion embedding plots for the layers of the implemented architecture. In the future, we plan to incorporate more modalities such as video and images and work on making the intersection matrix more informative.

5. References

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, no. 1, pp. 39–58, 2009.

[2] M. El Ayadi et al., "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] D. Amodei et al., "Deep Speech 2: End-to-end Speech Recognition in English and Mandarin," in *International Conference on Machine Learning (ICML)*, 2016, pp. 173–182.

[4] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end Speech Emotion Recognition using Deep Neural Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5089–5093.

[5] S. Sahoo, P. Kumar, B. Raman, and P. P. Roy, "A Segment Level Approach to Speech Emotion Recognition using Transfer Learning," in *Proceedings of the 5th Asian Conference on Pattern Recognition (ACPR)*, 2019, pp. 435–448.

[6] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 2594–2604.

[7] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," in *Conference on Artificial Intelligence (AAAI)*, vol. 33, no. 01, 2019, pp. 6818–6825.

[8] H. Feng, S. Ueno, and T. Kawahara, "End-to-end Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model," *InterSpeech 2020*, pp. 501–505, 2020.

[9] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal Speech Emotion Recognition using Pre-trained Language Models," *arXiv preprint arXiv:1912.02610*, 2019.

[10] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning Discriminative Features from Spectrograms using Center Loss for Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7405–7409.

[11] S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition using Audio and Text," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 112–118.

[12] S. Tripathi, S. Tripathi, and H. Beigi, "Multimodal Emotion Recognition on IEMOCAP Dataset using Deep Learning," *arXiv preprint arXiv:1804.05788*, 2018.

[13] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-tuning "BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition," *arXiv preprint arXiv:2008.06682*, 2020.

[14] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating Deep Learning Architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[15] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-Attention for Speech Emotion Recognition." in *InterSpeech*, 2019, pp. 2578–2582.

[16] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech Emotion Recognition using Multi-Hop Attention Mechanism," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2822–2826.

[17] R. Zhang, H. Wu, W. Li, D. Jiang, W. Zou, and X. Li, "Transformer based Unsupervised Pre-training for Aacoustic Representation Learning," *arXiv preprint arXiv:2007.14602*, 2020.

[18] M. Xu, F. Zhang, and S. U. Khan, "Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion," in *IEEE Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 1058–1064.

[19] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic Speech Emotion Recognition using RNNs with Local Attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227–2231.

[20] K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition using Deep Neural Network and Extreme Learning Machine," in *Fifteenth Annual Conference of the International Speech Communication Association (ISCA)*, 2014.

[21] U. Bhatt et al., "Explainable Machine Learning in Deployment," in *Conference on Fairness, Accountability, and Transparency*, 2020, pp. 648–657.

[22] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, "An Explainable Deep Fusion Network for Affect Recognition Using Physiological Signals," in *ACM International Conference on Information and Knowledge Management (CIKM)*, 2019, pp. 2069–2072.

[23] J. Zhang, X. Huang, L. Yang, and L. Nie, "Bridge the Semantic Gap between Pop Music Acoustic Feature and Emotion: Build an Interpretable Model," *Neurocomputing*, vol. 208, pp. 333–341, 2016.

[24] M. B. Fazi, "Beyond Human: Deep Learning, Explainability and Representation," *Theory, Culture & Society*, 2020.

[25] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," in *International Conference on Machine Learning (ICML)*, 2017, pp. 3145–3153.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated Feedback Recurrent Neural Networks," in *International Conference on Machine Learning (ICML)*, 2015, pp. 2067–2075.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[28] 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.

[29] R. Plutchik, "The Nature of Emotions," *Journal Storage (JSTOR) Digital Library's American scientist Journal*, vol. 89, no. 4, pp. 344–350, 2001.

[30] 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.

[31] 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.