INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multimodal Emotion Recognition using Cross-Modal Attention and 1D Convolutional Neural Networks

Krishna D N, Ankita Patil

HashCut Inc., India

krishna@sizzle.gg, ankita@sizzle.gg

Abstract

In this work, we propose a new approach for multimodal emotion recognition using cross-modal attention and raw waveform based convolutional neural networks. Our approach uses audio and text information to predict the emotion label. We use an audio encoder to process the raw audio waveform to extract high-level features from the audio, and we use text encoder to extract high-level semantic information from text. We use cross-modal attention where the features from audio encoder attend to the features from text encoder and vice versa. This helps in developing interaction between speech and text sequences to extract most relevant features for emotion recognition. Our experiments show that the proposed approach obtains the state of the art results on IEMOCAP dataset [1]. We obtain 1.9% absolute improvement in accuracy compared to the previous state of the art method [2]. Our proposed approach uses 1D convolutional neural network to process the raw waveform instead of spectrogram features. Our experiments also shows that processing raw waveform gives a 0.54% improvement over spectrogram based modal.

Index Terms: multimodal emotion recognition, cross-modal attention, 1D-CNNs.

1. Introduction

Speech emotion recognition is one of the key components for human-computer interaction systems. Humans express their emotions in various ways, including speech, facial expression, and so on. Studies have proved that emotion recognition using speech alone is ineffective for building emotion recognition models, but with extra guiding signals like face expression and the spoken content, we can get the emotion recognition working at a good accuracy level. In this paper, we study emotion recognition using audio and text.

Many algorithms have been proposed in the past for speech emotion recognition. Before the era of deep learning, many people proposed machine learning algorithms like Support vector machines, hidden markov models, Gaussian mixture models, and so on. Work by [4] propose to use HMM and SVM for speech emotion recognition. Work done by [5] uses GMMs for speech emotion recognition. Recent developments in deep learning techniques have shown to improve emotion recognition performance. Deep learning[3] has been successfully used by many fields in the speech area, including speech recognition [6], speaker recognition [7], speaker diarization [8], and so on. Most of the recent works on speech emotion recognition involve deep learning due to effectiveness in extracting very good high-level features from the audio, which helps in improving classification accuracy. Recently [10] propose to use deep neural networks to extract high-level features from audio and then uses extreme learning machines to predict class labels from these high-level features. Convolutional neural networks are very well known in the computer vision community due to their efficiency and effectiveness in extracting good features. Convolution neural network has also been used for speech emotion recognition by [11], and it is shown to perform well for SER task. Since speech is a temporal sequence, Recurrent neural networks are the best fit for processing speech signals. Work by [12] shows that using Recurrent neural networks(Bi-LSTM) is better for extracting high-level features and helps improving speech emotion recognition accuracy. The recent trend in Multimodal emotion recognition has shown that Multimodal emotion recognition performs better than unimodal emotion recognition models due to additional information provided by the second modality. Recent work in [13] uses both audio and transcript information for emotion recognition. Work done by [14] uses phoneme embeddings extracted from the text as a piece of extra information during classification. Recently [15] used speech embedding as an extract guiding signal along with speech features in a multimodal setting to improve emotion recognition accuracy. The recent development of attention models [9] has shown performance improvement in most of the sequence-related problems. Recently H. Xu et al. [2] shows that using an interactive attention model for multimodal emotion recognition can improve the emotion recognition system performance. In recent years, convolutional neural networks are shown to handle raw waveforms directly instead of handcrafted features[17,18,19]. These models can be trained end-to-end using back-propagation. Once these networks are trained with back-propagation, the model will extract very good emotion-specific features which help in improving systems performance. Many studies have proved that learning features from raw audio give better performance than using handcrafted features as input to the neural networks.

In this work, we propose cross-modal attention to learn interactive information between audio and text modality for multimodal emotion recognition. Our model has two networks called audio encoder and text encoder. The audio encoder takes raw audio waveform as input and applies 1D convolution operations, followed by LSTM, to obtain higher-level features from raw audio data. Similarly, the text encoder processes the word vectors using a 1D convolutional layer, followed by LSTM to obtain a higher-level feature sequence. Cross-modal attention is applied to the output sequences from audio encoder and text encoder, which helps in finding the interactive information between the audio and text sequences and thus helps in improving the performance. Our approach uses raw audio waveform as input to audio encoder instead of handcrafted features. Since audio encoder has 1D convolution operations, it learns features automatically during training. Our experiments show raw waveform processing gives better accuracy compared to spectrogram features. We conduct all the experiments on the IEMOCAP dataset. Specifically we use emotion classes which includes *angry*, *happy*,*sad* and *neutral* and we combine happy and excitement into happy class.

http://dx.doi.org/10.21437/Interspeech.2020-1190

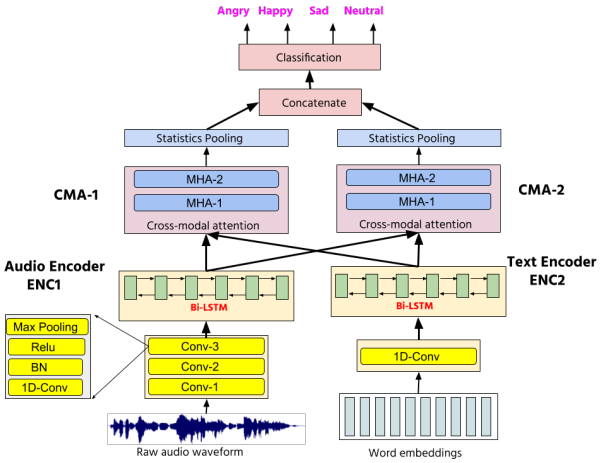

Figure 1: *Proposed model architecture.*

The organization of the paper is as follows. Section 2 explains our proposed approach in detail. In section 3, we give details of the dataset. In section 4, we explain our experimental setup in detail. Finally section 5 contains our results.

2. Proposed approach

In this paper, we propose a new architecture for multimodal emotion recognition. Our model uses both audio and text information to obtain better classification performance. The model architecture is shown in Figure 1. The proposed architecture consists of 2 streams, audio encoder (ENC1) and text encoder(ENC2) as shown in the figure. The audio encoder consists of a series of 1D convolutions followed by a Bi-LSTM layer to extract high-level feature representations from the raw audio waveform. Similarly, the text encoder (ENC2) takes sequence of Glove embeddings [20] for every sentences and feeds them a single 1D convolutional layer followed by a Bi-LSTM layer to obtain high-level semantic representations. These two feature sequences are fed into a cross-modal attention module in the audio encoder and text encoder, as shown in Figure 1. The cross-modal attention module in the audio encoder stream (CMA-1) takes the output from Bi-LSTM of the audio encoder as query vectors and output from the Bi-LSTM of the text encoder as key and value vectors and applies multi-head scaled dot product attetion. Similarly, The cross-modal attention module in the text encoder stream (CMA-2) takes the output from Bi-LSTM of the text encoder as query vectors and output from Bi-LSTM of the audio encoder as key and value vectors and applies multi-head scaled dot product attention. This helps in finding the interactive information between audio and text feature sequences. The CMA-1 and CMA-2 helps in selecting features which more relevant and helpful for emotion recognition. We finally pool the features from both CMA-1 and CMA-2 using statistics pooling layer. These pooled features are concatenated and fed to the softmax layer to predict the emotion label. We explain these blocks in detail in the following section.

2.1. Audio encoder

The audio encoder processes audio data in order to extract features that are useful for better emotion classification. The audio encoder consist of three 1D convolutional layers followed by a single Bi-LSTM layer. The audio encoder takes raw waveform as input and applies sequence of 1D convolution operations to extract meaningful features from the raw audio. Each convolutional block consist of a convolution operation followed by Batch normalization,Relu and max pooling operation as described in Figure 1. Each convolutional layer has its own kernel size and filters. we have [1x3,1x5,1x7] filter size for *Conv-1*, *Conv-2* and *Conv-3* layer respectively. Similarly, we have [64,100,100] filters for *Conv-1*, *Conv-2* and *Conv-3* layer respectively. Since convolutional layer does not capture temporal information, we use Bi-LSTM right after convolution. Let $\boldsymbol{X_n} = [\boldsymbol{x_1}, \boldsymbol{x_2}..\boldsymbol{x_n}, ...\boldsymbol{x_N}]$ be raw audio sequence with N samples.

$$\boldsymbol{F^A} = \text{Convolution}(\boldsymbol{X_n}) \tag{1}$$

Where, Convolution is a sequence of 3 1D convolutional layers applied to the raw waveform $\boldsymbol{X_n}$ as shown in Figure 1. After convolution, we obtain a feature sequence $\boldsymbol{F^A} = [\boldsymbol{f_1}, \boldsymbol{f_2}.....\boldsymbol{f_T}]$ of length T (T<<N). Typically after the convolution operation, $\boldsymbol{F^A}$ can be looked as a feature matrix whose x-axis is a time dimension, and the y-axis is a feature dimension. The feature dimension is the same as the number of filters in the last convolutional layer. In our case, the feature dimension is 100, as the number of filters in the last convolutional layer is 100. Since convolution can't capture temporal information, we use Bi-LSTM to process the feature vectors from the convolutional block.

$$\boldsymbol{H^A} = \text{Bi-LSTM}(\boldsymbol{F^A}) \tag{2}$$

Where Bi-LSTM represents a single bidirectional LSTM layer whose hidden dimension is 100. $\boldsymbol{H^A} = [\boldsymbol{h_1}, \boldsymbol{h_2}.....\boldsymbol{h_T}]$ represents the output sequence from the Bi-LSTM layer.

2.2. Text encoder

The text encoder model processes sequence of word embeddings extracted using the Glove word embedding model [20]. The text encoder(ENC2) consists of a single 1D convolutional layer followed by a single Bi-LSTM layer. Text encoder takes sequence of N-word embeddings $\boldsymbol{W^T} = [\boldsymbol{w_1}, \boldsymbol{w_2}.....\boldsymbol{w_N}]$ as input to 1D convolutional layer, where $\boldsymbol{w_i}$ is a 300 dimensional Glove word embedding for the $\boldsymbol{i}^{\text{th}}$ word. We set N to be 50 in our experiments. The convolutional layer acts as a projection layer where it projects the word embedding dimension from 300 to 100 and does not alter the time dimension.

$$\boldsymbol{F^T} = \text{Convolution}(\boldsymbol{W^T}) \tag{3}$$

Where $\boldsymbol{F^A} = [\boldsymbol{f_1}, \boldsymbol{f_2}.....\boldsymbol{f_N}]$ is a feature sequence after convolution operation. It can be noted that $\boldsymbol{W^T}$ and $\boldsymbol{F^T}$ has the same number of frames due to kernel size one during convolution operation. After convolutional layer, the text encoder passes feature sequence $\boldsymbol{F^T}$ to Bi-LSTM to capture high-level semantic and syntactic information.

$$\boldsymbol{H^T} = \text{Bi-LSTM}(\boldsymbol{F^T}) \tag{4}$$

Where Bi-LSTM represents a single bidirectional LSTM layer whose hidden dimension is 100. $\boldsymbol{H^T} = [\boldsymbol{h_1}, \boldsymbol{h_2}.....\boldsymbol{h_N}]$ represents the output sequence from the Bi-LSTM layer.

2.3. Cross modal attention

Attention is a very well known concept in deep learning, especially for temporal sequences. Attention models have shown state of the art results in various fields like, speech recognition,

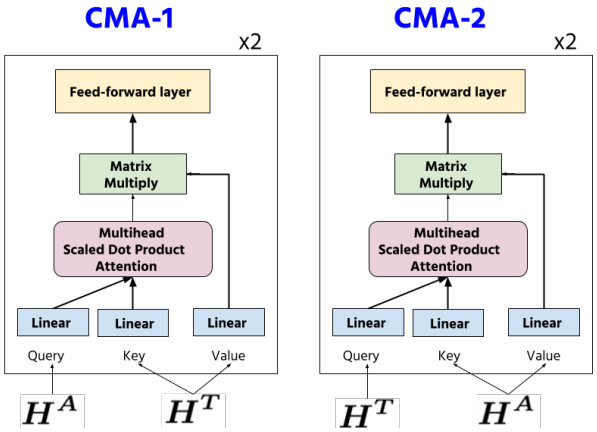

Figure 2: *Cross modal attention. CMA-1 (left) represents cross modal attention from audio to text and CMA-2(right) represents cross modal attention from text to audio*

speaker identification, language identification so and on. Various attention models are invented over the past, which includes dot product attention, locality-sensitive attention, additive attention, so on. In our case, we use a special type of attention called multi-head scaled dot product attention [22]. The scaled dot product attention is the core building block of cross-modal attention. The scaled dot product attention in multi-head settings are described in [22]. A detailed Cross-modal attention block is shown in Figure 2.

The cross-modal attention consists of Multi-head attention layers. The multi-head attention (MHA) layers consists of multi-head scaled dot product attention, matrix multiply, and position-wise feed-forward layer, as shown in Figure 2. The MHA block consists of M linear layers for query Key and Value matrices, where M is the number of heads in the multi-head attention. The MHA takes audio features and text features as input and applies attention logic to find the interactive information between the modalities. The inputs to the cross-modal attention blocks are query key and values, as shown in Figure 2. The cross-modal attention block takes audio feature or text features from Bi-LSTM and applies linear transform to create $\boldsymbol{Q_i}$, $\boldsymbol{K_i}$ and $\boldsymbol{V_i}$ using $\boldsymbol{i}^{\text{th}}$ linear transform where, $\boldsymbol{i} = [\boldsymbol{1, 2.....M}]$ and M is the total number of attention heads. The $\boldsymbol{Q_i}$, $\boldsymbol{K_i}$ and $\boldsymbol{V_i}$ are fed into scaled dot product attention layer followed by matrix multiplication between the value matrix and attention weights. The scaled dot product attention $\boldsymbol{A_i}$ for $\boldsymbol{i}^{\text{th}}$ head is defined as follows.

$$\boldsymbol{A_i} = \text{Softmax}(\frac{\boldsymbol{Q_i K_i}}{\boldsymbol{d_q}})\boldsymbol{V_i} \tag{5}$$

Where, $\boldsymbol{d_q}$ is the dimension of the query vector. We combine the attention output from all the heads using simple concatenation and feed them into the feed-forward layer.

$$\boldsymbol{A} = \text{Concat}(\boldsymbol{A_1, A_2, A_3 ... A_i .....A_M})W_0 \tag{6}$$

In our case, we have 2 cross-modal attention modules CMA-1 and CMA-2. The CMA-1 is responsible for finding interactive information from audio to text, and CMA-2 is responsible for computing the interactive information from text to audio. In the case of CMA-1, we consider $\boldsymbol{H^A}$ as a query matrix and $\boldsymbol{H^T}$ as the key and value matrix. Similarly, in the case of CMA-2, we consider $\boldsymbol{H^T}$ as a query matrix and $\boldsymbol{H^A}$ as the key and value matrix. This process makes sure the attention is applied in both directions to capture interactive features which are helpful for better emotion classification. We use layer normalization[20] before linear transformation during attention computation, which helps in faster convergence during training. Both CMA-1 and CMA-2 consists of 2 multi-head attention (MHA) blocks in our architecture. The audio feature matrix $\boldsymbol{H^A}$ consists of T vectors, each of dimension 100 and text feature $\boldsymbol{H^T}$ consist of N vectors of dimension 100.

2.4. Statistics pooling

The statistics pooling layer computes the mean and standard deviation across time in a feature sequence. Statistics pooling helps in pooling the features across time to obtain a single feature representation for classification. The mean and standard deviation are concatenated to obtain the second-order statistics of the feature sequence from CMA-1 or CMA-2. We finally concatenate second-order statistics from CMA-1 and CMA-2 for further classification.

$$\boldsymbol{P^A} = \text{Concat}(\text{mean}(\boldsymbol{A^a}), \text{std}(\boldsymbol{A^a})) \tag{7}$$

$$\boldsymbol{P^T} = \text{Concat}(\text{mean}(\boldsymbol{A^t}), \text{std}(\boldsymbol{A^t})) \tag{8}$$

Where, $\boldsymbol{A^a}$, $\boldsymbol{A^t}$ are the outputs from CMA-1 and CMA-2 respectively. $\boldsymbol{P^A}$ is a pooled feature vector from $\boldsymbol{A^a}$ and $\boldsymbol{P^T}$ is a pooled feature vector from $\boldsymbol{A^t}$. Both $\boldsymbol{P^T}$ and $\boldsymbol{P^T}$ are having same dimension.

3. Dataset

We conduct all our experiments on the IEMOCAP dataset[1]. IEMOCAP data is a publicly available research dataset for emotion recognition. IEMOCAP dataset contains information about face video along with the text and audio, and hence it is useful in multimodal emotion research. The dataset contains 12hrs of recording. The data is recorded during a conversation between 2 people and are manually annotated by human annotators. The dataset has ten speakers and five sessions. In each session, two person conversations is recorded, segmented into sentences, and labeled by professionals, and they are validated by three different evaluators. The dataset consists of a total of 5531 utterances from all five sessions. We use 4 emotions angry(1103), happy(1636), neutral(1708) and sad(1084). The happy data is a combination of happy and excited class. During training, we use leave-one-out session cross-validation technique as proposed in previous publications. We keep four sessions for training and test on the remaining session, and we repeat this procedure for all the five sessions. The final accuracy is the average of all test sessions.

4. Experiments

Our proposed approach consists of 2 streams, audio encoder (ENC1), text encoder (ENC2), and cross-modal attention(CMA-1 and CMA-2). The input to the audio encoder is a raw audio signal of 10sec duration. Since the sampling frequency of the audio data is 16KHz, we feed 160000 samples into the audio encoder. If the audio length is more than 10sec long, we crop the first 10sec. If the audio length is less than 10sec, we pad zeros. The audio encoder is a sequence of 3 convolutional layers with filer [1x3,1x5,1x7] and number of

filters [64,100,100]. Each of these convolutional layers consists of 1D convolution operation, 1D-batch normalization, and 1D max pooling. The initial convolutional layers of the audio encoder help in reducing the temporal dimension of the input data and also helps in extracting relevant features for emotion recognition. The output of the convolutional block from the audio encoder is fed to a Bi-LSTM with a hidden size of 100 and 0.2 dropout. Similarly, the text encoder has a single 1D convolutional layer with 100 filters, and the filter size is kept to 1. The input to the text encoder is a sequence of word embeddings extracted using Glove word embedding models. We assume maximum of 50 words in each sentence. The dimension of each word embedding is 300. The convolutional layer in the text encoder acts as a projection layer and the output of which goes to a Bi-LSTM with a hidden layer size of 100. We also use a dropout of 0.2 for the Bi-LSTM in the text encoder. The output of Bi-LSTM layers is 200 dimensions, as the hidden dimension is 100. We use projection layers to project this 200 dimension vectors into 100 dimensions. We use two layers of Multi-head attention for both CMA-1 and CMA-2. Each multi-head attention(MHA) layer uses ten attention heads and layer-normalization before scaled dot product operation. The forward feed layer inside MHA has a hidden dimension similar to its input, which, in our case, is 100. We use Adam optimizer [21] with a learning rate of 0.001. We use Pytorch [27] framework for implementation. All our models are trained on RTX 2080Ti Graphics cards.

5. Results

In this section, we compare our system with the recent state of the art system proposed for multimodal emotion recognition on IEMOCAP. Our system consists of an audio encoder that encodes raw audio using convolutional layers in order to extract meaningful feature representations from audio, and we use text encoder to extract high-level semantic information from word embeddings. We use cross-modal attention to compute the interactive information between the two modalities in order to improve system performance. It can be shown that our model outperforms the recently proposed state of the art system [2] by 1.9% absolute improvement in unweighted accuracy (Also known as class accuracy) as it can be seen in Table 1(Last row). Our method also shows that using raw waveform processing instead of handcrafted features can help in improving system performance. We show 0.52% absolute improvement in performance by using raw waveform processing instead of handcrafted features like spectrogram features, as it can be seen in Table 1(second last row). For this experiment, we used 257 dimensions spectral feature extracted for every 25ms window with a 10ms shift. We compute spectral features for 10sec audio and use it as input to the model.

Table 1: *Comparison of previous multimodal emotion recognition systems. Bold indicates the best performance*

System	Unweighted Accuracy
E-Vector [29]	57.25%
MCNN + LSTM [13]	64.33%
MCCN+phoneme embeddings [14]	68.50%
H.Xu et. al [2]	70.90%
CMA+spec (proposed)	72.24%
CMA+Raw waveform (proposed)	**72.82**%

Table 2: *Comparison of uni-modal emotion recognition models. Bold indicates the best performance*

System	Unweighted Accuracy
Audio-only	
TDNN+LSTM [18]	**60.70**%
LSTM+Attn [28]	58.80%
Self-Attn+LSTM(Ours)	55.60%
Text-only	
H.Xu et. al [2]	57.80%
Speech-Embedding [15]	60.40%
Self-Attn+LSTM	**65.90**%

We conduct experiments on different modalities to see the performance variations. We wanted to see our model performance on unimodal experiments, and we conduct audio-only and text-only experiments on the same dataset. For the audio-only experiment, we use the same audio encoder and Cross-modal attention and statistic pooling, but since we do not have text modality, we feed the audio features itself as query, key, and value vectors to cross-modal attention block. This setting is known as self attention in literature. Our audio-only experiment consist of a three convolutional layers, a single layer LSTM and a self attention layer. The performance of the audio-only experiment is show in Table 2(row 4). Similarly, for the text-only experiment, we feed text features itself as query key and value vectors to cross-modal attention block. Our text-only experiment consist of a single convolutional layer, a single layer LSTM and a self attention layer.The performance of text-only experiment model is show in Table 2(last row).

6. Conclusions

Speech emotion recognition is one of the most challenging and still unsolved problem in the speech community. In this work, we propose a new approach for multimodal emotion recognition using raw waveform based convolutional neural network in combination with cross-modal attention. Our approach uses raw audio processing using 1D convolutional models, and cross-modal attention networks between audio and text feature in order to obtain better emotion recognition system performance. Our approach shows that features extracted from raw audio and cross-modal attention mechanism can compute the interactive information between audio and text sequences, which are very helpful for emotion classification. Our experiments show that proposed architecture achieves the state of the art emotion classification accuracy on IEMOCAP dataset [1].

7. Acknowledgements

I want to thank HashCut Inc. for supporting this project. Any opinion, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of HashCut Inc., India.

8. References

[1] 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

[2] Xu H, Zhang H, Han K, Wang Y, Peng Y, Li X, "Learning Align-

ment for Multimodal Emotion Recognition from Speech", *Proc. Interspeech 2019*, 3569-3573

[3] Schmidhuber, Jürgen, "Deep learning in neural networks: An overview." Neural networks", 61 (2015): 85-117.

[4] Y.L.LI, G. Wei, "Speech emotion recognition based on HMM and SVM", *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Vol.8, 18-21 Aug. 2005, pp.4898 4901

[5] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," in *Ninth International Conference on Spoken Language Processing*, 2006

[6] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition", in *ICASSP*, 2016

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2018, April 2018, pp. 53295333.

[8] A. Zhang, Q. Wang, Z. Zhu, J. Paisley and C. Wang, "Fully Supervised Speaker Diarization", *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom*, 2019, pp. 6301-6305.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate", *arXiv preprint arXiv:1409.0473*, 2014

[10] 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*, 2014.

[11] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), New Orleans, LA, USA*, March 2017, pp. 51155119.

[12] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[13] Cho, J., Pappagari, R., Kulkarni, P., Villalba, J., Carmiel, Y., Dehak, N, "Deep Neural Networks for Emotion Recognition Combining Audio and Transcripts", *Proc. Interspeech* 2018.

[14] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram and phoneme embedding", *Proc. Interspeech 2018*, pp. 3688–3692, 2018.

[15] N, Krishna and Reddy, Sai, "Multi-Modal Speech Emotion Recognition Using Speech Embeddings and Audio Features", *AVSP 2019 Melbourne ,Autralia*

[16] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, , J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences", In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[17] M. Ravanelli and Y. Bengio, "Speaker Recognition from Raw Waveform with SincNet," *2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece*, 2018, pp. 1021-1028.

[18] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma,and N. Dehak, "Emotion identification from raw speech signals using DNNs," in *Proc. INTERSPEECH, Hyderabad, India*,2018, pp. 3097–3101

[19] K. D N, A. D, S. S. Reddy, A. Acharya, P. A. Garapati and T. B J, "Language Independent Gender Identification from Raw Waveform Using Multi-Scale Convolutional Neural Networks," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain*, 2020, pp. 6559-6563.

[20] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization", In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[23] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *IEEE SLT*, 2018.

[24] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[25] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), vol. 1, 2017, pp. 873–883.

[26] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "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)*. IEEE, 2016, pp. 5200–5204.

[27] Paszke, Adam and Gross, Sam and Chintala, Soumith and Chanan, Gregory and Yang, Edward and DeVito, Zachary and Lin, Zeming and Desmaison, Alban and Antiga, Luca and Lerer, "Adam: Automatic differentiation in PyTorch", in *NIPS*, 2017

[29] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE*, 2015, pp. 4749–4753