MULTIMODAL TRANSFORMER WITH LEARNABLE FRONTEND AND SELF ATTENTION FOR EMOTION RECOGNITION

Soumya Dutta and Sriram Ganapathy

Learning and Extraction of Acoustic Patterns (LEAP) lab, Indian Institute of Science, Bangalore.
soumyadutta@iisc.ac.in, sriramg@iisc.ac.in

ABSTRACT

In this work, we propose a novel approach for multi-modal emotion recognition from conversations using speech and text. The audio representations are learned jointly with a learnable audio front-end (LEAF) model feeding to a CNN based classifier. The text representations are derived from pre-trained bidirectional encoder representations from transformer (BERT) along with a gated recurrent network (GRU). Both the textual and audio representations are separately processed using a bidirectional GRU network with self-attention. Further, the multi-modal information extraction is achieved using a transformer that is input with the textual and audio embeddings at the utterance level. The experiments are performed on the IEMOCAP database, where we show that the proposed framework improves over the current state-of-the-art results under all the common test settings. This is primarily due to the improved emotion recognition performance achieved in the audio domain. Further, we also show that the model is more robust to textual errors caused by an automatic speech recognition (ASR) system.

Index Terms— Multi-modal emotion recognition, Transformer networks, self-attention models, learnable front-end.

1. INTRODUCTION

With the growing demand for conversational agents and personal assistants, automatic recognition of human emotion has emerged as a key task in enabling enhanced user experience. Human emotion recognition using multi-modal data of text, speech and video has substantial impact on various applications like smartphones, wearable devices, smart speakers, driver monitoring in auto-motives, mood analysis and mental health. This area of developing emotional intelligence would allow machines to be more human-like in the interactions [1].

The problem of emotion recognition is challenging primarily due to the complex process of emotion that is highly personal. The emotion in human interaction can be detected using facial expressions [2], speech [3], gestures [4] and physiological signals like respiration [5]. Further, different modalities contain varying degrees of information relating to emotion and hence, designing a joint multi-modal approach to emotion recognition is considered to improve the performance of these systems [6]. While the ability to perceive emotions in a multi-modal way is required, it is also necessary to perceive the emotions through each modality in a robust manner. In this paper, we explore an emotion recognition task with audio and text.

The advances in deep learning have also benefited the emotion recognition from speech [7], audio-visual emotion recognition [8] as well as from joint text, speech and visual modalities [9]. For multi-modal emotion recognition, logistic regression and support vector machine classifiers were explored by Sikka et al. [10] and Castellano et. al [11]. In these works, the datasets used were smaller (less than 500 samples). The recent works use larger datasets with deep learning methods. For example, the works by Yoon et al. [12] and Lee et. al. [13] focus on speech and text based emotion recognition, while Majumder et al. [14] explores emotion recognition using a tri-modal fusion of text, speech and video features. The log-mel filter bank features along with other acoustic indicators have been used extensively in the works related to emotion recognition in speech [15, 16, 17]. These features are knowledge driven, meaning that any variability in the dataset will not play an explicit role in the feature extracted from the audio files. This has motivated researchers to develop audio feature extractors which are learnable [18, 19, 20].

In this work, we propose an approach to emotion recognition in conversations, where we first extract learnable features from audio and text using LEAF [19] and BERT [21] respectively. We propose a method of information fusion across utterances with a self-attention network for each of the two modalities. The two modalities are combined in a multimodal transformer for superior classification using the long term multi-modal context information. Our model is evaluated on the widely used IEMOCAP dataset [22].

2. RELATED WORK

For audio emotion recognition, prosodic features have been investigated along with other acoustic features [23, 24]. Here, the mel frequency cepstral coefficients (MFCC) have been used along with pitch based features. The OpenSmile [25] is a widely used toolkit for extracting the audio features and has been used in several works like [15, 16]. Wu et. al [17] used a long term log-mel filter bank feature of 250ms frame length in addition to the pitch features.

The feature extraction from text has seen considerable changes in the past few years with the transfer learning ability of bidirectional encoder representations from transformers (BERT) [21]. This has been used for text feature extraction in [17]. In another work, the convolutional neural networks trained on word2vec representations [26] has been used for text based emotion recognition [15, 16]. The global vectors for word representations (GloVe) based embeddings [27] have also been applied for emotion recognition by Tripathi et. al. [28] and Pepino et. al [29].

The fusion of multi modal information had been achieved by using Bidirectional LSTM (Bi-LSTM) networks in [30]. This has been further improved upon by the Poria et. al. [15], where the fusion is performed in a hierarchical manner. The use of attention mechanisms for retaining the long term context has been explored in [12, 13]. The architecture of self attentive networks has also been applied in [17] for multi-modal emotion recognition.

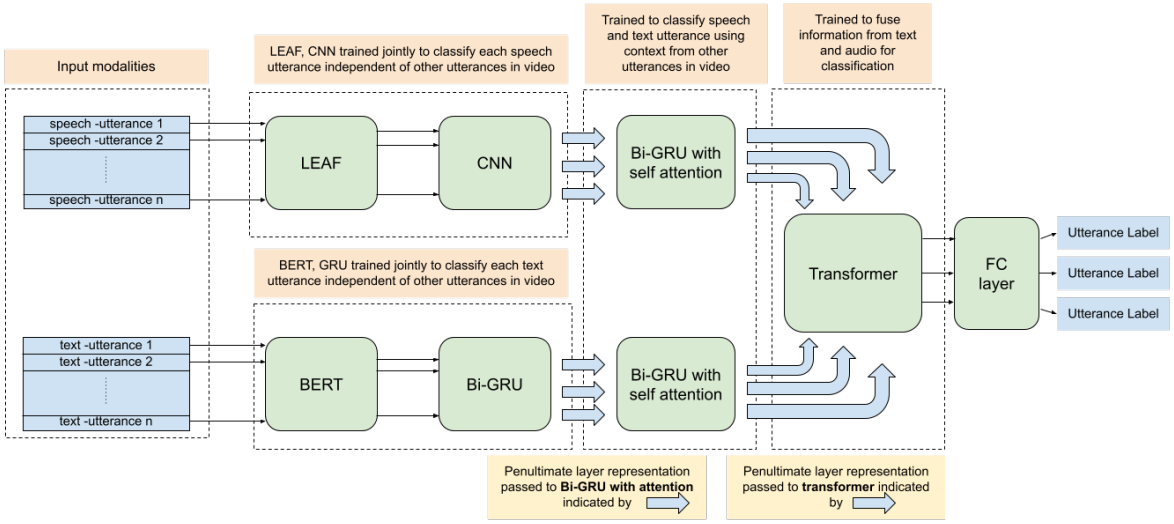

Fig. 1. Model Architecture - All blocks in green refer to trainable parts of the architecture.

3. PROPOSED APPROACH

A schematic of the major components of the model architecture is shown in Fig. 1.

3.1. Audio Feature Extraction

The audio features are extracted using the learnable front-end for audio classification (LEAF) [19] model, which learns the features using components like filtering, pooling, compression and normalization. Further, given the small footprint of the model, the learnable front-end can be integrated into the larger emotion recognition network and can be jointly learned. The LEAF model, employing 1-D convolutional networks on the raw audio files, generates discriminative spectrogram-like audio representations. The LEAF output is used in a CNN network with average pooling followed by a fully connected output layer. The LEAF-CNN network is jointly trained with utterance level labels and the 100 dimensional features from the subsequent fully connected layer are used as embeddings for each audio utterance. For the subsequent part of the model, this LEAF-CNN network is frozen, and the embeddings from this layer are used as inputs.

3.2. Text Feature Extraction

We use the BERT based features for text representations [21]. After each utterance text is passed through the BERT model, the last four hidden layers from the encoder are spliced together. This constitutes the BERT representations for each utterance. To add more context to the representations, the BERT outputs are passed through a 2 layer Bi-directional Gated Recurrent Unit (GRU) Model [31] with a hidden dimension of 100. The BERT-GRU model is jointly trained with utterance level labels and the representations at the output side of the GRU Model are considered as the utterance level text embeddings for the subsequent Bi-GRU with self-attention block. Like the LEAF-CNN model for audio, the BERT-GRU network is frozen for the subsequent parts of the model and the embeddings are used as inputs for the subsequent layers.

3.3. Multi-utterance self-attention

There is an emotion label for each utterance in the conversation. The objective then is to conditionally predict the output of embedding for utterance $u(t)$ using the embeddings from previous time instances, $u(t-1)$, $u(t-2)$ and so on along with future time instances such as $u(t+1)$, $u(t+2)$ and so on.

The addition of the contextual information proposed in this work is inspired by Poria et. al. [15]. The Bi-GRU network with attention is input with the LEAF-CNN utterance embeddings for the audio modality and the BERT-GRU utterance embeddings for the text. The objective of the attention network is to incorporate the context selectively from the previous and the future utterances to enable the prediction of the emotion label in the current utterance.

Let the dimension of the input to the self-attention layer from the Bi-GRU be $(B \times S \times H)$. Here, B represents the batch size, S represents the number of utterances in the conversation and $H = [H_f; H_b]$ is the concatenation of the forward (H_f) and backward (H_b) output dimensions at the utterance level. For simplicity, a batch size of 1 is assumed in these calculations. The matrix products will be replaced by the tensor product when $B > 1$.

Let $O_f \in \mathcal{R}^{S \times H_f}$ and $O_b \in \mathcal{R}^{S \times H_b}$ denote the outputs from the BiGRU that is input to the self-attention network. Let the two weight matrices, $W_f^a \in \mathcal{R}^{H_f \times H_f}$ and $W_b^a, \in \mathcal{R}^{H_b \times H_b}$ denote the attention layer parameters. The attention in the forward direction is computed as,

$$A_f = (O_f W_f^a)(O_f W_f^a)^T \quad (1)$$

$$A_f^{ij} = \frac{exp(A_f^{ij})}{\sum_{j=1}^{S} exp(A_f^{ij})} \; \forall i, j \in \{1, 2, \ldots, S\} \quad (2)$$

$$O_f^a = A_f O_f \quad (3)$$

The attention in the backward direction is identical. The Bi-GRU with self-attention block is trained to jointly predict the label of all the utterances in a video. This output, $[O_f^a; O_b^a]$, is used as the embeddings for each utterance for the subsequent multimodal transformer. Like the feature extractors, the Bi-GRU with self-attention network is frozen while training the subsequent multimodal transformer.

Model arch.	Modality	Acc. (%)
Bi-GRU w/o self attn.	Audio	69.2 ± 0.7
Bi-GRU + self attn.	Audio	73.8 ± 1.7
Bi-GRU w/o self attn.	Text	76.3 ± 0.9
Bi-GRU + self attn.	Text	79.1 ± 0.7
Hidden dim.	Modality	Acc. (%)
50	Audio	72.9 ± 1.9
100	Audio	$\mathbf{73.8 \pm 1.7}$
200	Audio	72.7 ± 2.6
50	Text	78.1 ± 0.4
100	Text	$\mathbf{79.1 \pm 0.7}$
200	Text	79.0 ± 0.5

Table 1. The results for different hidden dimension choices in Bi-GRU network as well as the experiments with and without attention.

3.4. Multi-modal Transformer

The Bi-GRU with attention generates utterance level representations for both the speech and text. These embeddings are spliced. The transformer encoder [32] is employed for the fusion of the audio and text information for improved emotion recognition. It is hypothesized that the transformer with self attention layers is able to handle long-term dependencies and to combine the modalities more effectively than the LSTM or GRU models. Our implementation of entire model is publicly available.[1]

4. EXPERIMENTS AND RESULTS

4.1. Dataset

The experiments reported in this work use the IEMOCAP [22] database. The dataset has a total of 151 video recordings divided into 5 sessions. Each session has a separate pair of a male and female actors conversing with each other. Every video is separated into individual utterances, all of which are labelled by human annotators as belonging to one of the 10 emotions - angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted or "other". However, keeping in line with previous works [15, 17], we have used 5 label categories in our task, namely, angry, happy, sad, neutral and excited. Further, the happy and excited classes are merged. Thus we have a total of 5531 utterances from the four emotion labels (happy 1636, angry 1103, sad 1084, neutral 1708). All the utterances have the transcriptions along with the audio files.

4.2. Choice of hyper-parameters

As the proposed model contains three components, namely, the feature extractors (BERT-GRU for text and LEAF-CNN for audio), the bidirectional GRU blocks and the transformer, choice of the hyper-parameters for each of them plays a role in the performance of the overall model. These parameters are chosen based on the performance of each of the three parts with Session 5 as the test set and Session 1 as the validation set. All the experiments reported in this work use 10 random weight initialization choices. The mean performance and the standard deviation using the random initializations are reported in all the experiments below.

The LEAF-CNN model for the audio classifier is memory intensive in its basic implementation. In order to circumvent this issue, the stride length of the LEAF features is increased to 30ms instead of the default value of 10ms. The LEAF-CNN feature extractor is trained with a batch size of 16 and a learning rate of 1×10^{-5}. The dimension of the audio representation from the CNN classifier was checked for 3 different values of 50, 100 and 200. Based on the performance on the validation data, the audio representation dimension from the LEAF-CNN classifier was fixed at 100.

Similarly, the BERT-GRU model for the text feature extractor is trained with a batch size of 32 and an identical learning rate. For this utterance level embedding extractor, the dimension of 200 is found to give the best accuracy on the validation set.

The Bi-GRU with self-attention, for both text and audio, is trained with a batch size of 32 and learning rate of 1×10^{-3}. The accuracy for both the modalities with hidden dimensions of 50, 100 and 200 as well as the configuration with and without self-attention in the BiGRU layers is shown in Table 1. The best performance is achieved for a hidden dimension of 100. The self attention in the Bi-GRU network was found to be essential for the improved performance of the overall model. This was consistent for both the text and the audio modality.

The final part of the model, namely the multi-modal transformer, is trained with a batch size of 32 and a learning rate of 1×10^{-4}. Several experiments are carried out with different combinations of number of hidden layers and hidden layer dimensions as shown in Table 2. The variation of the results with the number of attention heads was found to be negligible and hence it was fixed at 12. The final multi-modal transformer configuration is chosen to have a hidden layer dimension of 120, with 12 attention heads and 3 hidden layers.

# of layers	Hidden dim.	Acc (%)
2	60	83.6 ± 0.4
2	120	83.5 ± 0.5
2	240	83.8 ± 0.3
3	60	83.8 ± 0.7
3	120	$\mathbf{83.8 \pm 0.7}\%$
3	240	$83.2 \pm 0.6\%$
4	60	$83.8 \pm 0.6\%$
4	120	$83.6 \pm 0.7\%$
4	240	$83.0 \pm 0.7\%$

Table 2. The results with different hyper-parameter choices in the multi-modal transformer.

4.3. Evaluation setting

4.3.1. Five fold validation (CV-5)

In this setting, the models are trained on four sessions and are validated on the fifth. The average validation accuracy is calculated over the five sessions. It is to be noted that, since our model contains three parts, namely, feature extractors, Bi-GRU with self-attention models and the transformer, it is necessary to create 5 separate models for each of the three stages to avoid any leakage of test data in the training data. Further to alleviate the effects of randomness, our models are run for 10 different initializations and the average and standard deviation of our results are reported. The results on 5-fold setting are reported in Table 3.

4.3.2. Ten fold Cross Validation results (CV-10)

In this test setting, leave-one-speaker-out cross validation is performed, in which our models are trained on 9 speakers and tested

[1] https://github.com/iiscleap/multimodal_emotion_recognition

Setting	Modality	Acc. (%)
CV 5	Text	75.1 ± 0.2
CV 5	Audio	70.0 ± 0.3
CV 5	Text + Audio	$\mathbf{78.9 \pm 0.2}$
CV 10	Text	78.7 ± 0.2
CV 10	Audio	73.8 ± 0.4
CV 10	Text + Audio	$\mathbf{82.2 \pm 0.3}$
Session 5	Text	79.1 ± 0.7
Session 5	Audio	73.8 ± 1.7
Session 5	Text + Audio	$\mathbf{83.8 \pm 0.7}$

Table 3. The results for different evaluation settings for the proposed model.

System	Test Setting	Acc. (%)
Wu et al. [17]	CV-5	63.5
This work	CV-5	**74.9**
This work	CV-10	76.2
This work	Session 5	77.3

Table 4. The results of the emotion recognition systems evaluated with ASR transcripts instead of the reference text.

on the 10th. As in the case of five-fold cross validation, the 10 different models in this case are trained and tested. The results with this setting are reported in Table 3. The trends seen for the CV-5 condition is also similar to the CV-10 results.

4.3.3. Session 5 as test

The results with Session 5 as the test session are also provided. For this, one of the other 4 sessions is randomly chosen for validation (which in our case was Session 1). The results averaged over 10 different initializations are reported in Table 3.

4.4. Evaluation with ASR generated transcripts

Deep learning models using speech and text for emotion recognition are trained with the audio files along with their text transcriptions provided in datasets. In practice, we rarely encounter audio with transcriptions. This test setting, when a model is trained with provided transcripts and tested on ASR outputs, tests the robustness of the model to the noise in the text modality. For obtaining the ASR transcripts, we use the Google Speech to Text[2] on IEMOCAP audio files. The results under this test setting are shown in Table 4. The results show that, even with the ASR output, the performance of the model does not degrade drastically under all the three test settings. A considerable improvement is seen in the performance of our model over the results reported in [17] under the CV-5 testing strategy (relative improvement of 31%).

5. DISCUSSIONS

The results from other recent works on this dataset are mentioned in Table 5. Our proposed model achieves a better accuracy than the previous works in the literature under all the three test settings reported. While there are improvements on all evaluation settings, the performance on the held-out speaker (CV-10) improves the state-of-art results (Wu et al. [17]) by a relative margin of 18 %. Further, in the audio modality alone, the proposed approach yields an accuracy

[2] https://cloud.google.com/speech-to-text

System	Test Setting	Modality	Acc. (%)
Poria et al. [15]	Session 5	T, A, V	76.1
Wu et al. [17]	Session 5	T, A	83.2
This work	Session 5	T, A	**83.8**
Yoon et al. [33]	CV-5	T,A	71.8
Wu et al. [17]	CV-5	A	61.3
Wu et al. [17]	CV-5	T, A	78.4
This work	CV-5	T, A	**78.9**
Poria et al. [34]	CV-10	T, A, V	76.1
Yoon et al. [12]	CV-10	T, A	77.6
Li et al. [35]	CV-10	T,A	79.2
Wu et al. [17]	CV-10	T, A	78.3
This work	CV-10	T, A	**82.2**

Table 5. Comparison with other works under the three different test settings. We note that for [17] and [12], we compare our results with the unweighted accuracy.

of 70% in CV-5 condition (Table 3), while the state-of-art system on audio modality achieves an accuracy of 61.3% (Table 5). Thus, the major improvements from this work are primarily in the audio domain. The robustness of the model to noise in text modality, as shown in Table 4, further establishes the improvement in the performance using only the speech inputs.

The accuracy gain can be attributed to a number of factors in the proposed system. The learnable front-end features from LEAF enable the model to capture local time-frequency patterns of audio. The combination of GRU with attention in the Bi-GRU layer helps the model to describe the long-term information in the audio signal. Further, the multi-modal transformer helps the fusion of the text and audio modalities and allows the improvements observed in the audio domain to enhance the multi-modal emotion recognition performance.

6. CONCLUSION

In this paper, a hierarchical information fusion architecture has been proposed for conversational multimodal emotion recognition. Our proposed architecture improves the representation of utterance level speech and text at each stage of information fusion, first by jointly learning the representations and then, by employing self-attention over other utterances in the recording. A transformer is used for effective multimodal fusion of the two modalities. Our proposed model achieves state-of-the-art performance on the IEMOCAP dataset under all the three test settings reported in literature. Further, the model is shown to be robust to textual errors in ASR transcripts.

7. REFERENCES

[1] Maja Pantic, Nicu Sebe, Jeffrey F Cohn, and Thomas Huang, "Affective multimodal human-computer interaction," in *ACM international conference on Multimedia*, 2005, pp. 669–676.

[2] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn, "Face alignment through subspace constrained mean-shifts," in *2009 IEEE 12th ICCV*. IEEE, 2009, pp. 1034–1041.

[3] Klaus R Scherer, Tom Johnstone, and Gundrun Klasmeyer, *Vocal expression of emotion.*, Oxford University Press, 2003.

[4] Costanza Navarretta, "Individuality in communicative bodily behaviours," in *Cognitive Behavioural Systems*, pp. 417–423. Springer, 2012.

[5] R Benjamin Knapp, Jonghwa Kim, and Elisabeth André, "Physiological signals and their use in augmenting emotion recognition for human–machine interaction," in *Emotion-oriented systems*, pp. 133–159. Springer, 2011.

[6] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[7] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[8] Yelin Kim, Honglak Lee, and Emily Mower Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *ICASSP*. IEEE, 2013, pp. 3687–3691.

[9] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *AAAI*, 2020, vol. 34, pp. 1359–1367.

[10] Karan Sikka, Karmen Dykstra, et al., "Multiple kernel learning for emotion recognition in the wild," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 517–524.

[11] Ginevra Castellano, Loic Kessous, and George Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and emotion in human-computer interaction*, pp. 92–103. Springer, 2008.

[12] Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP*. IEEE, 2019, pp. 2822–2826.

[13] Woo Yong Choi, Kyu Ye Song, and Chan Woo Lee, "Convolutional attention networks for multimodal emotion recognition from speech and text data," in *Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML)*, 2018, pp. 28–34.

[14] Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-based systems*, vol. 161, pp. 124–133, 2018.

[15] Soujanya Poria, Erik Cambria, et al., "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)*, 2017, pp. 873–883.

[16] Zexu Pan, Zhaojie Luo, Jichen Yang, and Haizhou Li, "Multi-modal attention for speech emotion recognition," *arXiv preprint arXiv:2009.04107*, 2020.

[17] Wen Wu, Chao Zhang, and Philip C Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *ICASSP*. IEEE, 2021, pp. 6269–6273.

[18] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *SLT*. IEEE, 2018, pp. 1021–1028.

[19] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi, "LEAF: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.

[20] Debottam Dutta, Purvi Agrawal, and Sriram Ganapathy, "A multi-head relevance weighting framework for learning raw waveform audio representations," *arXiv preprint arXiv:2107.14793*, 2021.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[22] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[23] Iker Luengo, Eva Navas, Inmaculada Hernáez, and Jon Sánchez, "Automatic emotion recognition using prosodic parameters," in *Interspeech*, 2005.

[24] Shashidhar G Koolagudi, Nitin Kumar, and K Sreenivasa Rao, "Speech emotion recognition using segmental level prosodic analysis," in *2011 international conference on devices and communications (ICDeCom)*. IEEE, 2011, pp. 1–5.

[25] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[26] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin, "Advances in pre-training distributed word representations," *arXiv preprint arXiv:1712.09405*, 2017.

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[28] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," *arXiv preprint arXiv:1804.05788*, 2018.

[29] Leonardo Pepino, Pablo Riera, Luciana Ferrer, and Agustín Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *ICASSP*. IEEE, 2020, pp. 6484–6488.

[30] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

[31] Kyunghyun Cho, Van Merriënboer, et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[33] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, "Multimodal speech emotion recognition using audio and text," in *SLT*. IEEE, 2018, pp. 112–118.

[34] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[35] Runnan Li, Zhiyong Wu, Jia Jia, Yaohua Bu, Sheng Zhao, and Helen Meng, "Towards discriminative representation learning for speech emotion recognition.," in *IJCAI*, 2019, pp. 5060–5066.