MULTIMODAL SPEECH EMOTION RECOGNITION USING TRANSFORMER BASED FUSION

NIKHIL RAJU SHINDE
210897232
NIKI MARIA FOTEINOPOULOU
MSC ARTIFICIAL INTELLIGENCE

Abstract— Humans possess the cognitive ability to comprehend and assimilate information derived from diverse sources, encompassing oral communication, written language, and visual representations. The advancement of deep learning technology has led to a notable enhancement in the precision of speech recognition. The identification of emotions from speech is a vital characteristic, and the utilization of deep learning technology has significantly enhanced the precision and responsiveness of emotion recognition. There remain a multitude of hurdles in the pursuit of enhancing precision. This study aims to examine different neural networks and fusion techniques to improve the precision of emotion recognition. In this study, we provide a fusion model based on transformers that aims to acquire significant multimodal information pertaining to the presentation of emotions from both speech and text modalities. The experimental findings conducted on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset demonstrate that our suggested strategy attains a classification accuracy of 70%.

1. INTRODUCTION

The automatic detection of emotion can be achieved through various modalities, such as face expressions, speech patterns, and bodily movements. In situations where visual cues such as face expressions are not available, voice serves as the sole modality for recognizing emotions. This is particularly evident in contexts such as telephone conversations, audio messaging, and call center applications (Petrushin, 1999). Efficient and informed feedback can be provided promptly by implementing an automated system that can accurately detect the emotional state of the caller. Nevertheless, a significant drawback of several speech emotion recognition algorithms is their inadequate performance, primarily attributed to their exclusive reliance on speech characteristics. This research presents a proposal for enhancing the performance of Speech Emotion Recognition through the integration of speech and text information. The aforementioned concept is derived from the observation that written content can be extracted from spoken language, hence facilitating the process of emotion recognition. For instance, an individual engaged in conversation can discern emotional cues not alone through auditory perception, but also through interpreting the semantic content of spoken expressions. Furthermore, individuals have a tendency to employ particular vocabulary choices in order to convey their emotions during verbal communication, as they have acquired knowledge regarding the association between certain words and the corresponding emotional states (C. M. Lee, 2002). The existing literature on pattern recognition indicates that the incorporation of multimodal information leads to enhanced performance in comparison to utilizing a single modality (Tripathi, 2018). The incorporation of speech and text data in Speech Emotion Recognition has the potential to enhance its performance. This improvement is driven by the need to achieve more genuine human-computer interaction through the recognition of expressiveness in speech. Furthermore, several technologies can use these advancements to derive benefits in this domain.

This study presents a straightforward approach to enhance the precision of Speech Emotion Recognition through the integration of speech and text characteristics. In order to assess the effectiveness of the suggested approach, which involves utilizing a mix of speech features and Bert-word embeddings, we conducted an evaluation by performing speech emotion recognition solely based on speech features, and text emotion recognition. The term "speech feature" is used in this study to refer to an array of speech features that are derived from the spoken components of an utterance after the pre-processing phase of eliminating silence. Through the utilization of transformer-based fusion (Singh, 2023), the integration of voice and text features is anticipated to surpass the utmost performance achieved by either speech or text feature independently.

2. Related work

The research of Speech Emotion Recognition has gained significant importance in recent years. This section presents an overview of Speech Emotion Recognition (SER) approaches that utilize features extracted from Automatic Speech Recognition (ASR) systems. Additionally, we discuss strategies for mitigating the limitations imposed by ASR faults on the performance of SER. Recent research has focused on the investigation of approaches for Speech Emotion Recognition that utilize features extracted from Automatic Speech Recognition. These methods have demonstrated notable advancements in the field, achieving state-of-the-art performances in Speech Emotion Recognition. Several studies have explored methods for integrating speech and text data derived from Automatic Speech Recognition outcomes (S. Yoon, May 2019), (Shin, 2019), (W. Wu, 2021), [28], [29]. Nevertheless, achieving appropriate fusion of auditory and text elements may provide a significant challenge.

In their study, (C. M. Lee, 2002) suggested employing a combination of text and speech features using a logical "OR" function at the decision level to integrate information from both speech and language domains. In their study, Qin Jin et al. suggest the integration of auditory and lexical data, which are subsequently trained using a Support Vector Machine (SVM) classifier for the purpose of emotion category recognition. (Q. Jin, 2015)

Previous approaches employ transfer learning from a pre-trained Automatic Speech Recognition model (Beigi, 2020), (N. Tits, 2018), or employ the intermediate layer output of Automatic Speech Recognition as a feature for Speech Emotion Recognition (Zhang, 2022). Nevertheless, the data obtained through these methodologies lacks the inclusion of textual information, which is crucial for the identification and interpretation of emotional cues. Furthermore, a study conducted by researchers (Y. Li, 2022) demonstrates that incorporating text features derived from Automatic Speech Recognition outcomes, along with other input features, yields superior performance in Speech Emotion Recognition compared to methods that do not utilize text information. This finding underscores the significance of text information in enhancing the accuracy of Speech Emotion Recognition.

Several earlier studies (S. Yoon, May 2019) (Shin, 2019) (W. Wu, 2021) have examined the efficacy of Speech Emotion Recognition techniques on transcriptions and Automatic Speech Recognition outcomes. The empirical findings indicate that the utilizations of Automatic Speech Recognition outcomes lead to a decline in the overall performance of Speech Emotion Recognition when contrasted with the utilizations of transcriptions. The impact of emotions on the Automatic Speech Recognition has been a persistent concern that has garnered attention in numerous research investigations.

Alswaidan and Menai (2020) offered an extensive taxonomy of ways for recognizing emotions in text, along with a comprehensive elucidation of key elements and various approaches. (Nourah Alswaidan, 2020). The survey provides an explanation of numerous explicit and implicit techniques used for emotion recognition in text. It discusses many approaches discovered in the literature, highlighting their primary features, advantages, limitations, and providing comparisons. However, it fails to provide an analysis or discussion of the specific technologies and datasets utilized.

Abdullah et al. (2021) introduced many approaches for multimodal emotion detection, which involve the integration of different modalities, including images and text, as well as the combination of facial expressions with physiological responses of the body. The results obtained from this survey demonstrate that the integration of different modalities in the process of emotion recognition leads to enhanced accuracy in identifying and categorizing emotions. (Sharmeen M.Saleem Abdullah Abdullah, 2021).

3. Dataset & Features

3.1 Dataset

The accuracy of the proposed approach is evaluated using the IEMOCAP dataset, which is widely recognized as a benchmark dataset for emotion recognition. The IEMOCAP dataset was created during a time when traditional machine learning techniques, including SVM(Support-Vector Machine),logistic regression,Decision Trees and early neural networks, were the predominant approaches employed for Speech Emotion Recognition. Furthermore, it is employed for the assessment of contemporary

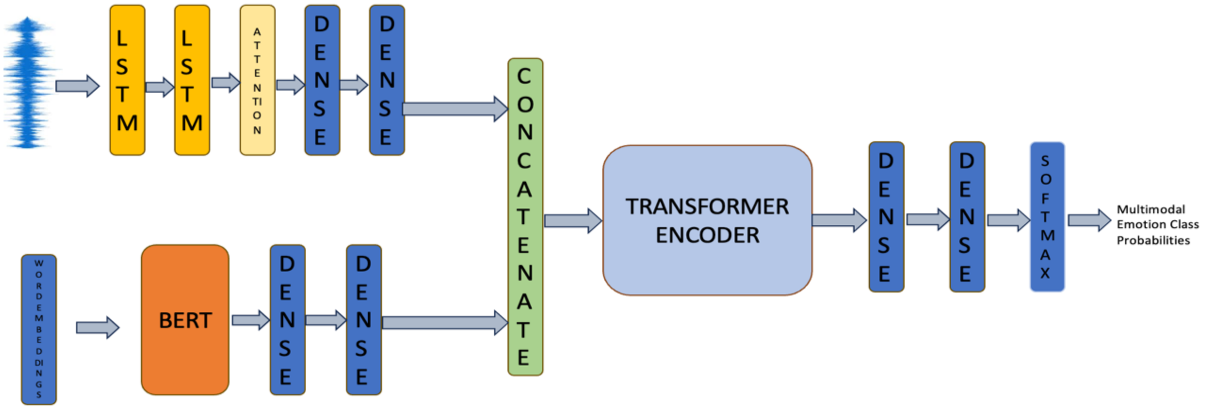

Fig.1 Proposed method depicting the Speech Emotion Recognition and Text Emotion Recognition models and the Transformer encoder fusion approach.

methodologies based on deep learning. The IEMOCAP dataset comprises around 12 hours of recorded speech. The IEMOCAP dataset comprises both scripted and improvised emotional speeches that have been categorized into five sessions. Each session features a male and a female speaker. The IEMOCAP dataset consists of ten speakers, evenly divided between males and females, with five speakers of each gender. The data utilized in our study consisted of four distinct emotion classifications, namely happy, sad, neutral, and angry.

TABLE 1. Given dataset specifications.

Dataset	IMEOCAP
Speakers	5 males & 5 Females
Utterance duration	1s-19s
No of classes	Happy - 1689
	Sad - 1084
	Neutral - 1708
	Angry - 1103

3.2 Audio -preprocessing

In order to derive characteristics from vocal segments, the speech files within the dataset are initially interpreted as vectors. Silence is eliminated from each individual utterance, as per file, in order to generate speech segments. Feature extraction is performed on the acquired speech segments for each utterance in our study. The application of a Hamming window involves segmenting each vocal utterance into frames and subsequently shifting them by overlapping steps. A comprehensive set of 34 features is extracted for every frame, encompassing various domains such as MFCCs, chromas, time domain, and spectral domain. These features include 13 MFCCs, 13 chromas, and 3-time domain features. Additionally, spectral domain features are considered, along with time domain features. In this study, we utilize 100 windows/segments, resulting in a feature size of (100, 34), which is then used as input for the model.

The algorithm utilized for eliminating silent segments from the speech data involves the following steps:

1. To define silence, a predetermined loudness threshold and minimum duration are established. The threshold was set at 0.04 for this study, and the minimum duration was 100 milliseconds.
2. Sample by sample, the voice signal is scanned. A sample's amplitude is counted as the beginning of a prospective silent segment when it drops below the predetermined threshold.
3. The counting continues if the amplitude remains below the threshold. Once the amplitude exceeds the threshold again, the counting stops.
4. If the total number of counted samples is greater than the minimum duration, that segment is considered silence and removed from the speech signal.
5. This process is repeated for the entire signal, segmenting out all regions that meet the silence criteria based on the set threshold and minimum duration values.

3.3 Text preprocessing

From the manual transcription, we extract the text of each speech and store it in a variable. To get words from a single utterance, we tokenized each syllable and then constructed into a sequence that is padded with up to 537 tokens. We make use of pre-trained GloVe Embeddings for CNN and LSTM models. With GloVe, any word in a corpus of text can be intuitively converted into a position in a three-dimensional space. As a result, terms that are similar will be grouped together. We insert the word embeddings matrix from pre-training into an embedding layer. (Jeffrey Pennington, 2014)

We use embeddings produced by the BERT tokenizer for the BERT language model. In order to best match our linguistic data, this tokenizer greedily generates a fixed-size vocabulary comprising individual letters, subwords, and words(McCormick, 2019).This vocabulary includes four items:

- Whole words
- Subword units that appear at the beginning of a word or in isolation, such as the subword "em" in the word "embeddings," are assigned the same vector representation as the freestanding sequence of characters "em" in the phrase "go get em." On the other hand, subword units that do not appear at the beginning of a word are denoted by the prefix "##" to indicate this particular circumstance.
- Words that are not located at the beginning of a word and are preceded by the symbol '##' to indicate this condition.
- Individual characters.

In order to do word tokenization according to this approach, the tokenizer initially verifies whether the entire term is present in the vocabulary. If the term cannot be found in the vocabulary, the system attempts to divide it into the largest subwords that are present in the vocabulary. If this is not possible, the system will resort to breaking down the word into individual characters.

4. Methodology

Multimodal learning is a specialized domain within the science of artificial intelligence that aims to efficiently process and analyze input originating from diverse modalities. In essence, this entails combining data from many sources, including text, images, audio, and video, in order to construct a more comprehensive and precise comprehension of the underlying information. Multimodal learning techniques facilitate the efficient processing and analysis of data from several modalities, hence enhancing the comprehensiveness and precision of the derived insights.

4.1 Speech-based emotion recognition:

The computational cost of analyzing entire speeches and the inclusion of extraneous information in all utterances might lead to suboptimal performance. One potential approach to address the aforementioned issue is utilizing the segmented speech segment of the utterance and eliminating periods of silence in order to extract relevant features. While this concept is not novel, a majority of research papers have employed complete speech utterances for the purpose of feature extraction, as demonstrated in (Tripathi, 2018). The utilization of voice-only segments for speech identification has been subject to criticism and is not employed by certain researchers. They contend that silence serves as an effective cue for emotion recognition (H. M. Fayek, 2017). In this study, speech properties are derived from the speech segment, taking into consideration the aforementioned benefits.

This study evaluates three different types of speech emotion recognition models. The aforementioned systems are: The concept of a "dense model" refers to a computational model that contains a high concentration of parameters or features. 2) Long Short-Term Memory (LSTM) networks. The utilizations of Long Short-Term Memory (LSTM) networks in conjunction with attention models. The attention model is employed to selectively extract significant information from the preceding layer through the utilizations of attention weights, hence enhancing the quality of language translation. The attention-based model demonstrated its superiority over alternative methodologies in the context of voice recognition. Specifically, the model exhibited exceptional performance in the task of mapping spoken words to the written domain. Drawing inspiration from the achievements of attention models in the domains of machine translation and voice recognition, we anticipate a comparable enhancement in the field of speech emotion recognition, given the similarity of the tasks at hand.

The first architecture included four fully connected layers, each consisting of 1024, 512, 256, and 4 units, respectively. The second architectural design involves the stacking of two LSTM layers. First layer has 512 hidden units which return the hidden state output for each input time step and the second layer has 256 units. Two FC(fully-connected) layers are subsequently appended to the neural network architecture, having 256 and 4 units respectively. The third model consists of the LSTM layers that has 2,041,348 trainable parameters. In initial layer, the utilization of LSTM is observed followed by an attention decoder of hidden units 256 and 128 respectively fully connected layers are included. Both dense layers in the model employ the Rectified Linear Unit (ReLU) and SoftMax activation functions.

4.2 Text-based emotion recognition:

For the text-based classifier, we experimented with three distinct architectures. One of the techniques employed is the use of Convolutional Neural Networks (CNN) featuring a series of four convolution layers of 1-Dimension with kernel size of 3 and a dropout rate of 0.2 (equivalent to 20%). Following the embedding layer, the convolution layers are used of units 256,128,64 and 32. Additionally, all layers utilize the Rectified Linear Unit (ReLU) activation function. In the final layer, a dense layer of 4 units is incorporated. The network under consideration possesses a cumulative count of trainable parameters amounting to 5,264,288.

The second network comprises of a two stacked Long Short-Term Memory (LSTM) architecture. The first layer and the second layer is composed of 512 and 256 hidden units respectively. Following the inclusion of the second layer, two FC(fully connected) layers are introduced, each consisting of 128 units with ReLU activation function and 4 units with SoftMax activation function respectively. The network possesses 3,410,288 trainable parameters.

The third network utilized in this study is a pretrained model known as Bidirectional Encoder Representations from Transformers (BERT). In the present study, the base model employed comprises of 12 transformer blocks, each with a hidden size of 768 and 12 attention heads. Fig 2 shows BERT's architecture. Each encoder block encapsulates a more sophisticated model architecture. A one-dimensional max-pooling layer is incorporated, which is then followed by a fully linked layer consisting of 128 units and a dropout rate of 0.2 (equivalent to 20%). In the final layer, a dense layer is incorporated, consisting of four units that correspond to the entire number of emotion categories.

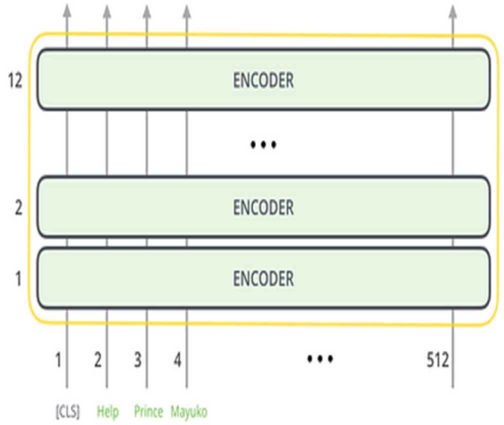

Fig 3. BERT's architecture

4.3 Multimodal based model:

We tested three techniques of fusion for the multi-modal model, including: 1) Feature concatenation Two approaches that have been proposed for fusion in attention-based models are attention-based fusion and transformer-based fusion.

The initial approach involves constructing a text model comprising of a sequence of four convolution layers of 1-Dimension, which are subsequently followed by a dense layer. The layers are composed of 256, 128, 64, 128, and 128 units, in that order. In the speech model, two Long Short-Term Memory (LSTM) layers are sequentially layered, and they are then followed by a dense layer. The dense layer consists of 512, 256, and 128 units, respectively. The speech and text networks are combined by concatenating them, and then sent through two FC(fully connected) layers of 128 with ReLU activation function and 4 units with SoftMax activation function respectively.

In the second approach, we employed the identical text and speech model as described in the first approach, and afterwards incorporated an Attention layer from the Keras library for the purpose of fusion. (Stephan Baier, 2017) The attention layer evaluates the significance of the representations inside the concatenated vector. Next, the architecture consists of two completely connected layers, each of 128 and 4 units, respectively. These layers employ the ReLU and SoftMax activation functions.

In the third model, we incorporated a Transformer encoder for the purpose of fusion. The inputs of the encoder initially pass via a self-attention layer, which enables the encoder to examine other elements within the input vector while encoding a particular unit from the vector. The outputs generated by the self-attention layer are subsequently passed to a feed-forward neural network. In our study, we employed a configuration consisting of eight attention heads and a feed forward network of 64 units. Multiple versions of a unimodal model were experimented with for the purpose of Transformer-based fusion. We tried different variations of unimodal model for Transformer based fusion.

A) CNN+LSTM

B) BERT + LSTM (Attention Decoder)

C)BERT (frozen layers) + LSTM (frozen layers).

In model 3C, the term "frozen layers" pertains to layers that remain unaltered in terms of their weights and biases throughout the training process.

4.4 Training and Inference

The objective function to be reduced throughout the training phase is defined as the summation of Lr1, Lr2, and Le. In the process of inference, the SoftMax function is utilized to transform the logits array of emotion classes generated by the model into probabilities for each class. The final determination of the classification outcome is made by selecting the class that possesses the highest probability.

The training and testing process were accomplished utilizing the GPU in the studies. Due to the substantial computational requirements entailed in the training phase of our suggested methodologies, it is advisable to employ a Graphics Processing Unit (GPU) for the purpose of training. The experimental setup employed in this study consists of four NVIDIA GeForce RTX 2070 graphics processing units (GPUs), each equipped with 32 GB of random-access memory (RAM). Additionally, an Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz central processing unit (CPU) with 12 GB of RAM is utilized. All models are developed using the TensorFlow framework. In each training experiment, the Adam optimizer is utilized with a learning rate of 10-4, and the default exponential decay rate is employed for the moment estimations.

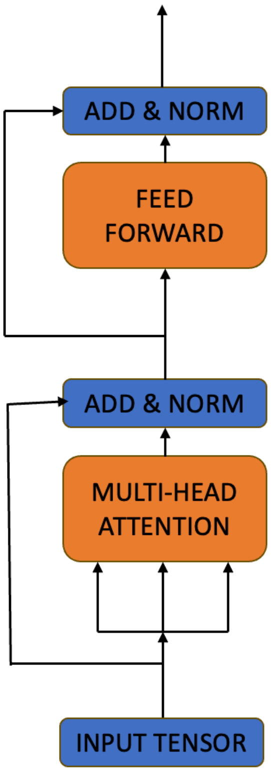

Fig 3.Transformer Encoder

Table 2. Accuracy result of Emotion Recognition using Speech Models

MODEL	ACCURACY
Dense model	42.42%
LSTM model	48%
LSTM + Attention Decoder model	51%

Table 3. Accuracy result of Emotion Recognition using Text Models

MODEL	ACCURACY
CNN model	63.66%
LSTM model	63.19%
BERT	66%

5. RESULTS

The results in Table 2 indicate considerable enhancements in accuracy when employing more sophisticated models for both speech and text modalities. The speech models exhibited an improvement in accuracy, rising from 42.42% when employing a dense neural network to 51% when using LSTM with an attention decoder. This highlights the advantages of applying Long Short-Term Memory (LSTM) in the context of modeling sequential speech data, as well as the attention mechanism's ability to recognize emotionally significant segments within the speech signal.

Table 4. Accuracy results of Multimodal Emotion Recognition Models

Text Model	Speech Model	Fusion Method	Accuracy
CNN	LSTM	feature concatenation	62%
CNN	LSTM	Transformer based fusion	64%
Bert	LSTM	Attention based fusion	48%
Bert	LSTM+Attention	Transformer based fusion	67%
BERT (frozen layers)	LSTM (frozen layers)	Transformer based fusion	70%

Table 3 presents the results pertaining to the accuracy of text emotion recognition achieved by the utilization of Glove word embedding input and BERT word embeddings. The findings indicate that the BERT model achieves the highest accuracy rate of 66%, which can be attributed to its utilization of a larger number of trainable parameters. The outcomes of comparing CNN and LSTM models exhibit a degree of similarity, albeit with LSTM demonstrating a reduced number of trainable parameters in comparison to CNN. Based on our analysis, it can be inferred that LSTM outperforms CNN in the context of text emotion recognition when utilizing Glove embeddings as input.

The outcome of the integration of speech and text features is presented in Table 4. In accordance with the preceding section, an assessment is conducted on five models. The optimal outcome is attained by employing model 3C, which incorporates the BERT network for textual input, LSTM networks for speech input, and a Transformer Encoder for fusion. The fusion network achieves an accuracy rate of 70% while utilizing a relatively small number of trainable parameters, specifically 5,213,060. The integration of speech and text information has the potential to enhance the efficacy of emotion recognition from speech, given their inherent relationship and derivability from speech data.

6. CONCLUSION

The present work conducted an examination of speech emotion recognition by utilizing voice features and word embeddings as inputs for several classifier designs. The proposed methodology employed a pre-trained BERT model for extracting textual features, while an LSTM model was applied for extracting speech features. The merging of these properties was achieved by employing a Transformer encoder. The findings of this study indicate that the utilization of a multimodal method yielded a 70% accuracy rate when applied to the IEMOCAP dataset, surpassing the performance of unimodal voice and text models.

The results underscore the advantages of integrating speech and text modalities in order to enhance the accuracy of emotion recognition, as opposed to relying solely on either mode in isolation. The utilization of transformer-based fusion proved to be efficacious in acquiring integrated representations and significant interplays among the modalities. This suggests that there is more emotional information present in both speech and text, which can be utilized to enhance performance.

Although the model has achieved a certain level of accuracy, there are still areas where improvements can be made. Potential avenues for further investigation may encompass the utilization of diverse pre-trained language models, the refinement of hyperparameter tuning processes, and the exploration of more intricate fusion methodologies. Conducting experiments on supplementary datasets would further substantiate the applicability of the proposed methodology across many areas. In general, the study effectively showcased the capacity of multimodal learning to enhance emotion identification capabilities, bringing them closer to achieving a level of comprehension comparable to that of humans.

7. ACKNOWLEGEMENT

I would like to express my sincere gratitude to Professor Niki Maria Foteinopoulou for their invaluable guidance, mentorship, and support throughout the course of this research project. Niki Maria Foteinopoulou provided continuous encouragement and valuable insights that enriched the quality of this research. Their willingness to devote time to in-depth discussions and their thoughtful feedback significantly contributed to the refinement of the research methodology and analysis.

8.Works Cited

Anon., n.d. s.l.:s.n.
Beigi, S. Z. a. H., 2020. A transfer learning method for speech emotion recognition from automatic speech recognition.

C. M. Lee, S. S. N. L. A. a. R. P., 2002. Combining Speech and Language Information for Emotion Recognition,.

Jeffrey Pennington, R. S. C. M., 2014. GloVe: Global Vectors for Word Representation.
Martin Wollmer, F. W. T. K. B. S. C. S. K. S., n.d. Youtube movie reviews: Sentiment analysis in an audio-visual context. *2013.*

McCormick, C., 2019. *BERT Word Embeddings Tutorial.* [Online]
Available at:
https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/

N. Tits, K. E. H. a. T. D., 2018. ASR-based features for emotion recognition: A transfer learning approach.

Nourah Alswaidan, M. E. B. M., 2020. A survey of state-of-the-art approaches for emotion recognition in text.

Petrushin, V., 1999. *"Emotion in speech: Recognition and application to call centers.".* s.l.:s.n.

Q. Jin, C. L. S. C. a. H. W., 2015. Speech emotion recognition with acoustic and lexical features,.

S. Yoon, S. B. S. D. a. K. J., May 2019. Speech emotion recognition using multi-hop attention mechanism.

Sharmeen M.Saleem Abdullah Abdullah, S. Y. A. A. M. A. M. S. S. Z., 2021. Multimodal Emotion Recognition using Deep Learning.

Shin, E. K. a. J. W., 2019. 'DNN-based emotion recognition based on bottleneck acoustic features and lexical features.

Singh, A., 2023. Transformer-Based Sensor Fusion for Autonomous Driving: A Survey.

Stephan Baier, S. S. V. T., 2017. Attention-based Information Fusion using Multi-Encoder-Decoder Recurrent Neural Networks.

Tripathi, S. a. H. B., 2018. Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning.

W. Wu, C. Z. a. P. C. W., 2021. Emotion recognition by fusing time synchronous and time asynchronous representations.

Y. Li, P. B. a. C. L., 2022. Fusing ASR outputs in joint training for speech emotion recognition.

Zhang, C. C. a. P., 2022. Channel and temporal-wise attention RNN leveraging pre-trained ASR embeddings for speech emotion recognition.