SPEECH EMOTION RECOGNITION USING MULTI-TASK LEARNING AND A MULTIMODAL DYNAMIC FUSION NETWORK

Sreyan Ghosh, S Ramaneswaran, Harshvardhan Srivastava, S. Umesh

1University of Maryland, College Park, USA, 2 IIT Delhi, India, 3Nvidia, Bangalore, India, 4Speech Lab, Department of Electrical Engineering, IIT Madras, Chennai, India

ABSTRACT

Emotion Recognition (ER) aims to classify human utterances into different emotion categories. Based on early-fusion and self-attention-based multimodal interaction between text and acoustic modalities, in this paper, we propose MMER, a multimodal multitask learning approach for ER from individual utterances in isolation. Our proposed MMER leverages a multimodal dynamic fusion network that adds minimal parameters over an existing speech encoder to leverage the semantic and syntactic properties hidden in text. Experiments on the IEMOCAP benchmark show that our proposed model achieves state-of-the-art performance. In addition, strong baselines and ablation studies prove the effectiveness of our proposed approach. We make our code publicly available on GitHub.

Index Terms—emotion recognition, computational paralinguistics, multimodal

1. INTRODUCTION

Humans express emotions implicitly in conversations, in addition to the explicit message that they convey. The task of Emotion Recognition (ER) aims at identifying these implicit emotions, which proves to be one of the key components of a better human-computer interaction systems.

Humans tend to express emotions in a variety of ways including body language, facial expressions, choice of words, tone of voice, and more. A variety of systems have been proposed for Speech Emotion Recognition (SER), including systems that exploit spectral, prosodic, or voice quality-based features in speech. With the recent advancement of deep learning-based end-to-end systems, high-level features learned through deep neural networks like CNNs or RNNs have outperformed such hand-engineered features. For speech specifically, the use of features learned through self-supervision has shown great success in the recent past.

Though in a conversational setting with spoken utterances speech might provide some of the most important signals for identifying the emotion in an utterance, the main challenge behind using speech as the sole modality for ER is that both low-level and high-level features easily overfit to noise or signals irrelevant to the task. Adding to this, we acknowledge that unimodal speech systems might find it difficult to correctly classify the emotion of an utterance in a conversation with a more natural setting where unlike acted scenes or improvisations certain important prosodic and acoustic cues might be absent. Prior-work has also shown that linguistic information from text is better suited for valence recognition in SER.

In recent work, researchers have successfully exploited emotion signals present in the modalities of vision and text through deep architectures and self-supervised features obtained from representation models. Though exploiting high-level self-supervised speech representations have shown great success in the past for the task of SER in a unimodal learning setup, no existing work that fuse representations from multiple modalities, propose the fusion of text and speech self-supervised representations, trained independently in individual modalities to solve modality-specific tasks, for multimodal SER, with the only exception being [13]. [13] also make an important observation that unimodal speech models suffer a problem of prediction bias to the angry class which only alleviates on adding linguistic cues from text. Thus efficiently fusing self-supervised representations from individual modalities for multimodal SER remains a challenge (see multimodal baseline in Section 4.2 and 4.4) and this remains a primary objective of our work in this paper.

In this work, we propose a novel architecture to capture fine-grained multimodal emotional information from acoustic and text modalities by the use of both speech and its corresponding text transcript for SER. To effectively utilize both speech and text data, the model needs to jointly learn features from different domains, and the performance of such a learning approach largely depends on how well the model is able to capture inter-modality interactions and alleviate unimodal biases. Although some studies combined both features and trained a multimodal model (discussed in detail in Section 2.2), very few works in the past has focused on the temporal relationship between speech and text at a fine-grained level. We believe that, since the modalities of speech and text inherently co-exist in the temporal dimension,
a multimodal system will benefit from using the alignment information. To achieve this, we first adopt the use of powerful pre-trained contextualized representation models for both the text and speech modalities separately, where we use the RoBERTaBase architecture [14] as the text encoder and wav2vec-2.0 [15] as the speech encoder. Second, to better capture the implicit alignments between words and speech frames, we propose the use of a unique multimodal interaction module (MMI). MMI essentially couples the standard Transformer layer with a cross-modal attention mechanism to produce a speech-aware word representation and a word-aware speech representation for each input word. Next, we concatenate the utterance representation obtained from the speech encoder module and the MMI module to classify the emotion expressed behind each utterance. Finally, to largely eliminate the bias of the textual context and to better take into account the natural monotonic alignment between the acoustic modality and the textual modality, we propose to solve automatic speech recognition as an auxiliary task. Our MMI module adds minimal extra model parameters over a wav2vec-2.0 speech encoder and achieves state-of-the-art results on the IEMOCAP benchmark. To summarize, the main contributions of this paper are as follows:

- We propose a novel end-to-end multimodal multi-task learning (MTL) framework for speech emotion recognition. We propose to use a novel multimodal interaction module to capture the inter-modality dynamics between words and speech frames. Through experiments, we show that our learning framework achieves state-of-the-art SER results on the standard IEMOCAP [16] dataset on the 5-fold cross-validation setup.

- Through ablation study and strong unimodal baselines we verify the effectiveness of each modality, our unique multimodal interaction module, and the MTL approach.

2. RELATED WORK

The task of Emotion Recognition (ER) aims at identifying emotional states from different signals expressed by a human at each turn during a conversation. ER as a downstream task has been thoroughly studied in the past, including systems that consider each individual utterance separately or take into consideration the context of the conversation. In this section, we will discuss primarily the former, including unimodal and multimodal approaches for the same.

2.1. Unimodal Emotion Recognition

Speech as a modality is one of the most studied in the task of ER and is commonly known as SER. There is a considerable amount of literature on SER. Early research focused on extracting low-level features like Mel-frequency cepstral coefficients (MFCCs) and Filter Banks (FBanks) or hand-engineered features like speaker rate, voice quality, etc. These features were then fed to machine learning classifiers and which proved to perform relatively well in terms of classification accuracy. Thanks to deep learning, deep neural networks have achieved a considerable boost in SER performance and can handle raw waveforms or low-level features directly without the need for hand-engineered features [17] [18].

With recent advances in self-supervised learning (SSL), pre-trained SSL features, like Natural Language Processing (NLP) [19], have achieved state-of-the-art (SOTA) performance in various downstream speech processing tasks like Automatic Speech Recognition (ASR), Phone Recognition (PER), Speaker Identification (SID), etc. A comprehensive study can be found here [7]. The current state-of-the-art on SER [20] also uses wav2vec-2.0 as the speech encoder and solves the SER task together with ASR as an auxiliary task by minimizing the CTC loss of the network. A recent study also reveals how supervised MTL on SSL pre-trained features can help benefit the performance of a downstream task when the auxiliary task is chosen properly. Inspired by this, we also choose ASR as our auxiliary task for improving representations learned by our model for our final task of ER.

Other unimodal approaches include using just text [21] or features extracted from facial expressions [22] for ER in human conversations.

2.2. Multimodal Emotion Recognition

For multimodal approaches, the most common combination of modalities includes speech and text. Early studies in this area focused on late fusion of multimodal representations [10] [23] [24]. Though this technique is simple and is effective at modeling modality-specific interactions, it is not effective at modelling cross-modal interactions [25]. Early fusion to capture inter-modality interactions has also been explored [26]. However, in general, early fusion also suppresses modality-specific interactions and does not outperform late fusion methods in emotion recognition [25] [27]. For better modeling the interactions between modalities, researchers have proposed cross-modal attention (CMA) mechanisms [28] [29] [11] [13]. With CMA, features from one modality are allowed to attend to the other, and the interaction between the sequences from the two modalities enables the system to extract the most useful features for emotion recognition. While [28] [29] use dot-product attention, with recent advances in NLP and the rise of self-attention, the use of self-attention-based CMA mechanisms for ER has been gaining traction [30] [11] [13]. The architecture we propose in this paper uses stacked CMA modules as an integral component which results in a unique multimodal interaction framework for the modalities of speech and text.
3. PROPOSED METHODOLOGY

3.1. Problem Formulation

Suppose we have a dataset $D$ with $N$ utterances $\{u_1, u_2, u_3, \cdots, u_N\}$ and its corresponding labels $\{y_1, y_2, y_3, \cdots, y_N\}$. Here we assume each utterance $u_i$ has both speech cues $a_i$ and text cues $t_i$ available where $u_i \in (a_i, t_i)$. $t_i$ can be ASR transcripts or human-annotated transcripts. We formulate the task of ER as assigning an emotion label $y_i$ to each utterance $u_i$, where $y_i$ denotes the probability distribution that the utterance $u_i$ belongs to one of the $J$ unique emotions being studied in the dataset.

3.2. Feature Encoder

3.2.1. Contextualized Speech Representations

For encoding speech to obtain high-level contextualized representations, we use a pre-trained wav2vec-2.0 [15] as our raw waveform encoder. Wav2vec-2.0 is based on the transformer architecture and pre-trained in a self-supervised fashion by solving a contrastive task and minimizing the InfoNCE loss [31]. We use the pre-trained checkpoint released by Facebook, pre-trained on 960 hours of Librispeech, and use the wav2vec-2.0-base architecture for all our experiments. Wav2vec-2.0 outputs $J$ hidden states and we denote the $j$th hidden state or contextualized embedding from the raw audio input $a_i$ of utterance $u_i$ as $e_j^{a_i}$ where $e_j^{a_i} \in \mathbb{R}^{768}$. $J$ depends on the length of the raw audio file and the CNN feature extraction layer of wav2vec-2.0, which extracts frames from the raw audio with a stride of 20ms and hop size of 25ms. In our experiments, we always keep the CNN feature extraction layer parameters frozen and train only the self-attention-based context encoder while E2E SER fine-tuning.

3.2.2. Contextualized Word Representations

We use RoBERTa BASE from the transformers family as our contextualized text encoder to encode the transcript of the utterance and obtain rich contextualized token representations. RoBERTa as a text encoder has been used extensively in ER literature. For each input transcript, we first tokenize the input sentence and add extra starting and ending tokens, $<$s $>$ and $<$ /s $>$ respectively. For a total of $M$ tokens, the $m$th contextualized embedding for each token in text transcript $t_i$, of utterance $u_i$, is denoted by $e_m^{t_i} \in \mathbb{R}^{768}$. We use RoBERTa only as a feature extractor and do not train it while fine-tuning our model.

3.3. Convolution Sub-sampling

Before passing our contextualized speech representations for further processing, we subsample $J$ to $J = J/2$ by passing $e^{a_i}$ through a convolutional 2D subsampling operation. This operation consists of two 2D convolution layers where each is followed by a $\text{relu}$ activation function. We do this for 2 main reasons: 1) In self-attention the memory requirement scales quadratically with the input length for long input sequences, and 2) The subsampling of speech frames produces better attention maps for the task of ASR in self-attention models [16], thereby allowing models to focus on useful speech information and neglecting noise hidden in redundant speech frames.

3.4. Multimodal Interaction Module

Our Multimodal Interaction Module (MMI) consists of 3 Cross Modal Encoder (CME) blocks annotated as $B$, $C$, and $D$ in Fig. 1. Each of these 3 CME blocks are constructed similar to a generic transformer layer [32], where each layer is composed of an $h$-head CMA module [33], residual connections and feed-forward layers. In this section, we discuss in detail the purpose and working of each of the 3 CME blocks and the acoustic gate $E$.

3.4.1. Speech-Aware Word Representations

As shown in Fig. 1 to learn better token representations with the guidance of the associated spoken utterance, we feed wav2vec-2.0 embeddings $A \in \mathbb{R}^{d \times J}$ as queries and RoBERTa embeddings $T \in \mathbb{R}^{d \times M}$ as keys and values into CMA module of CME block $B$ as follows:

$$\text{CMA}(A, T) = \text{softmax} \left( \frac{W_q A^\top W_k T}{\sqrt{d/m}} \right) \left[ W_v, T \right]^\top$$

(1)

where $\{W_q, W_k, W_v\} \in \mathbb{R}^{d/m \times h}$ denote the query, key and value weight matrices respectively for the $i$th attention head. The final output representation of the CME block $B$ is now $P = (p_0, P_1, \cdots, P_{m-1})$.

Next, to address the fact that each generated representation $p_i$ in the previous block correspond to the $i$th acoustic embedding and not the token embedding, we feed $P$ to another CME block $C$, which treats the original RoBERTa embeddings $T$ as queries and $P$ as keys and values. Finally, we now obtain the final Speech-Aware Word Representations as $R = (r_0, r_1, \cdots, r_{j-1})$.

3.4.2. Word-Aware Speech Representations

To obtain the word-aware speech representations and align each word to its closely related frame or wav2vec-2.0 embeddings we make use of another CME block $D$ by treating $T$ as queries and $A$ as keys and values. The final representations obtained from the block can be denoted as $Q = (q_0, q_1, \cdots, q_{j-1})$. Phoneme alignment has been long studied in speech science and acoustics and we hypothesize that this step is important so that each word can assign relative importance to the frames or embeddings important or not important to it.
3.4.3. Acoustic Gate

Speech frames might encode redundant information like random noise and other redundant speech cues. Thus it is important to implement an acoustic gate $g$ which can dynamically control the contribution of each speech frame embedding. Following previous work, we implement an acoustic gate $g$ as follows:

$$ g = \sigma \left( W_g^T [R, Q] + B_g \right) $$

where $W_g \in \mathbb{R}^{2d \times d}$ is a weight matrix, $B_g \in \mathbb{R}^d$ is the bias, and $\sigma$ is the element-wise sigmoid function. Finally, based on the gate output, the final word-aware speech representations are obtained by $Q = g \cdot Q$.

Post this step, we concatenate the speech-aware word representations and word-aware speech representations to obtain our final cross-modal MMI representations $M \in \mathbb{R}^{2d}$ where $M = [Q ; R]$ and pass it through a linear transformation $I(.)$ which down-projects $M$ to again a $d$ dimensional space.

3.5. CTC Layer, Emotion Prediction Layer, and Multitask Learning

As mentioned earlier, we train our model under an MTL framework where we solve two tasks in parallel, ASR and ER. For ASR we calculate the CTC loss and for ER we calculate CrossEntropy loss.

For calculating the CrossEntropy with ground-truth emotion labels, we first employ max pooling $\text{mp}(.)$ over wav2vec-2.0 speech encoder ($A$) and MMI module ($M$) independently across the time-step axis and then concatenate the embeddings to obtain a single final embedding $\mathbb{R}^{2d}$. This final embedding is then passed through a linear transformation and softmax activation function as follows:

$$ \hat{y} = \text{softmax} \left( W_p^T \text{mp}(A); \text{mp}(M) \right) + B_p $$

where $\hat{y} \in \mathbb{R}^4$ is the single vector representation for each utterance, $W_p \in \mathbb{R}^{2d \times 4}$ is a weight matrix, $\text{softmax}(.)$ denotes the softmax activation function, and $\text{mp}(.)$ denotes the attention pooling operation across the embedding axis. $\text{ap}(M)$ is further projected to $\mathbb{R}^{d}$ from $\mathbb{R}^{2d}$ using a linear layer, before concatenation and $\text{softmax}(.)$. Post this step, CrossEntropy is calculated by $L_{CE} = \text{CrossEntropy}(\hat{y}, y_i)$. Next, to calculate the CTC loss, we first pass the raw unpooled embeddings $A$ from the wav2vec-2.0 encoder through a linear layer as follows:

$$ \hat{i} = \text{softmax} \left( W_c^T A + B_c \right) $$

where $\hat{i} \in \mathbb{R}^{J \times V}$ and $J$ is the number of speech frames output by the wav2vec-2.0 CNN feature extractor and $V$ is the size of our vocabulary or the number of unique characters and symbols in our corpus and an extra blank token. $W_c \in \mathbb{R}^{d \times V}$ and $B_c$ is the added bias. Post this step, we calculate the CTC loss by $L_{CTC} = \text{CTC}(\hat{i}, t)$, where $t \in \{ t_0, \ldots, t_i, \ldots, t_N \}$ is a pre-processed version of the original $t_i$ where we remove all punctuation and convert all characters to uppercase.

Finally we minimize the sum of the two, weighted by a hyper-parameter $\alpha$ like $L = L_{CE} + \alpha L_{CTC}$. During inference, to obtain the final emotion label $y_i$ for utterance $u_i$, we simply drop the CTC linear layer and perform $\text{argmax}(\hat{y})$.

4. EXPERIMENTS

4.1. Dataset

Following much of the prior art in SER literature, we train and evaluate all our models on the IEMOCAP dataset [16]. IEMOCAP contains approximately 12 hours of speech from a total of 10 speakers, all of which comes from 5 scripted sessions, acted by professional actors. To keep our dataset settings consistent to prior-art and for a fair comparison, we evaluate our models on utterances assigned to one of the five emotions: Happy, Angry, Neutral, Sad and Excited and merge all samples labelled with Excited to Happy. For evaluation, we follow the five-fold cross-validation approach, where at each fold we leave one session out as the test set and take the average of the weighted accuracy obtained at each fold.
4.2. Baselines and Compared Methods

We build unimodal baselines with just text and speech modalities, where the text baseline uses RoBERTa$_{BASE}$ as the contextualized text encoder followed by a single linear layer and softmax activation for classification. For the unimodal speech baseline, we use an exactly similar setup but replace our encoder with pre-trained wav2vec-2.0-base, pre-trained on 960hrs of LibriSpeech [34]. Additionally, we also build a naive multimodal baseline where we simply concatenate pooled self-supervised representations $\text{mp}(A)$ and $\text{mp}(T)$ in a single-task SER learning setup.

We compare our model with other methods in literature evaluated on either of the 5-fold or 10-fold cross-validation setups, including unimodal and multimodal approaches. All results for prior art have been taken from literature (weighted accuracy unless stated otherwise). We only re-implement the current state-of-the-art approach [35] under the 5-fold cross-validation setup for a fair comparison.

4.3. Experimental Setup

We use the PyTorch framework to build, train and evaluate all our models. All pre-trained text and speech encoders were downloaded from the Huggingface library. Since we use the $base$ architectures for both RoBERTa and wav2vec-2.0, our $d$ effectively takes a value of 768. We trained and evaluated all our models with a batch size of 2 and accum-grad of 4 for a total of 100 epochs. For training, we kept the learning rate constant at $1e^{-5}$, which worked well for all our setups. For our multi-task learning setup, we trained our models with $\alpha \in \{0.1, 0.01, 0.001\}$ (analysis can be seen in Table 2). All optimal hyper-parameters were found via grid-search.

4.4. Experimental Results

In Table 1 we report the average weighted accuracy of our MMER model, averaged across 5-folds, compared against all our benchmarks and prior-art. As we see, MMER achieves SOTA on the IEMOCAP benchmark, with the closest being [9] where the author uses 2 contextualized speech encoders resulting in more than double the number of parameters as ours. We also see a 1% drop in WA for [35] when re-implemented on the 5-fold CV setup. Adding to these, both [9, 35] are unimodal setups and don’t leverage text cues which might benefit in various settings as discussed in Section 1. MMER also benefits from minimal trainable parameter addition over [35] or a simple wav2vec-2.0. We achieve 75.0% WA when google transcripts were used instead of gold transcripts for inference.

5. ABLATION STUDY

In this section, we conduct detailed analysis to highlight the key design choices of our proposed MMER.

### Table 1: Emotion Recognition Results on IEMOCAP

| Method                  | CV  | Modality | WA  |
|-------------------------|-----|----------|-----|
| **Unimodal Prior-art**  |     |          |     |
| Wu et al. [36]          | 10-fold | {a}   | 72.7% |
| Sajjad et al. [37]      | 5-fold  | {a}   | 72.3% |
| Lu et al. [38]          | 10-fold | {a}   | 72.6% |
| Liu et al. [39]         | 5-fold  | {a}   | 70.8% |
| Wang et al. [40]        | 5-fold  | {a}   | 73.3% |
| Pappagari et al. [41]   | 5-fold  | {a}   | 70.3% |
| Peng et al. [42]        | 5-fold  | {a}   | 62.6% |
| **Multimodal Prior-art**|     |          |     |
| Morais et al. [9]       | 5-fold  | {a,t} | 77.4% |
| Chen et al. [43]        | 5-fold  | {a,t} | 74.3% |
| Padi et al. [44]        | 5-fold  | {a,t} | 75.0% |
| Makiuchi et al. [45]    | 5-fold  | {a,t} | 73.5% |
| Chen et al. [13]        | 5-fold  | {a,t} | 74.3% |
| Cai et al. * [35]       | 10-fold | {a,t} | 78.1% |
| **Baselines**           |     |          |     |
| RoBERTa$_{BASE}$        | 5-fold  | {t}   | 69.2% |
| wav2vec-2.0             | 5-fold  | {a}   | 73.9% |
| multimodal              | 5-fold  | {a,t} | 74.1% |
| Cai et al. * (Ours)     | 5-fold  | {a,t} | 77.1% |
| **Proposed**            |     |          |     |
| MMER ($\alpha = 0.1$)† | 5-fold  | {a,1} | 78.1% |

5.1. Effect of CTC Loss

In this subsection, we aim to study the effect of multi-task learning on SER with CTC-based fine-tuning for ASR as the second task. As mentioned earlier, the hyper-parameter $\alpha$ is used to control the contribution of CTC loss. Therefore, we trained our model with $\alpha \in \{0.1, 0.01, 0.001\}$. As we see in Table 1 we achieve the best performance with $\alpha = 0.1$, and we observe a drop in performance of about 0.7% on the removal of the auxiliary CTC task ($\alpha = 0$).

### Table 2: Effect of $\alpha$

| $\alpha$     | WA  |
|--------------|-----|
| 0            | 76.9% |
| 0.1          | **78.1%** |
| 0.01         | 76.7% |
| 0.001        | 76.7% |

### Table 3: Effect of pooling

| Pooling Operation | WA  |
|-------------------|-----|
| max               | 78.1% |
| mean              | 77.6% |
| attention         | 77.7% |
| stats             | 77.5% |
| mean+max          | 77.9% |

CTC fine-tuning helps wav2vec-2.0 representations improve both word meaning and word identity information [46] as seen in Fig. 2 which also explains why the best performance was obtained with the highest CTC weight. Fig 2 also shows layer-wise CKA scores with F-Bank features, where we see CTC fine-tuning does not improve acoustic content of wav2vec-2.0 representations.
5.2. Effect of Pooling Operation

As mentioned in sub-section 5.2, we employ max pooling as our choice of pooling for obtaining the final emotion embeddings from $A$ and $M$. We also experimented with other pooling strategies, including mean, max, mean+max, and statistics pooling. Table 3 shows the results of all these setups where everything is kept constant except the pooling operation. As we see, max pooling achieves the best results for our setup. On the contrary, statistics pooling shows the least performance, which we hypothesize might be due to implicitly retaining speaker information.

5.3. Bias

Fig. 3 shows the confusion matrices for all 3 baselines and our best MMER model. Though MMER makes the most correct predictions for the Neutral emotion, we notice a lot of Happy emotions wrongly labeled as Neutral. Additionally, MMER improves overall baselines in bias towards the Angry emotion, which was noticed in prior-art.

5.4. Effect of Speech Representation

In this section, we analyze how different types of wav2vec-2.0 embeddings $A$ affect our final SER performance. Wav2vec-2.0 has a total of 12 encoder layers and one extra convolutional feature extractor where the layer learns different properties associated with the final downstream task. Optionally, the encoder layers can be frozen during task-specific fine-tuning. To be precise, we don’t change the input to the CTC layer but only change the input to our MMI module and the final pooling operation. We also change our final pooling operation to mean pooling to be consistent with prior art.

6. CONCLUSION

In this paper, we propose MMER, a novel multi-task multimodal approach for SER from spoken utterances using a dynamic multimodal fusion network. MMER outperforms all other unimodal and multimodal approaches in literature. As part of future work, we would like to investigate newer architectures and better auxiliary tasks that improve performance on SER benchmarks.
7. REFERENCES

[1] Wenjing Zhu and Xiang Li, “Speech emotion recognition with global-aware fusion on multi-scale feature representation,” in IEEE ICASSP 2022, 2022, pp. 6437–6441.

[2] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, “Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders,” in IEEE ICASSP 2020, pp. 6419–6423.

[3] Sreyan Ghosh, Samden Lepcha, S Sakshi, Rajiv Ratn Shah, and S Umesh, “Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances,” Interspeech 2022.

[4] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, “Meld: A multimodal multi-party dataset for emotion recognition in conversations,” arXiv preprint arXiv:1810.02508, 2018.

[5] Bernhard Kratzwald, Suzana Ilić, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger, “Deep learning for affective computing: Text-based emotion recognition in decision support,” Decision Support Systems, vol. 115, pp. 24–35, 2018.

[6] Stappen et al., “The muse 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress,” in Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, 2021, pp. 5–14.

[7] Yang et al., “SUPERB: Speech Processing Universal PERformance Benchmark,” in Interspeech 2021, 2021, pp. 1194–1198.

[8] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, “Emotion recognition from speech using wav2vec 2.0 embeddings,” arXiv preprint arXiv:2104.03502, 2021.

[9] Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz, “Speech emotion recognition using self-supervised features,” arXiv preprint arXiv:2202.03896, 2022.

[10] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi, “Multi-modal emotion recognition on iemo-cap dataset using deep learning,” arXiv preprint arXiv:1804.05788, 2018.

[11] DN Krishna and Ankita Patil, “Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks,” in Interspeech 2020, 2020, pp. 4243–4247.

[12] Puneet Kumar, Vishesh Kaushik, and Balasubramanian Raman, “Towards the explainability of multimodal speech emotion recognition,” in Interspeech 2021, pp. 1748–1752.

[13] Weidong Chen, Xiaofeng Xing, Xiangmin Xu, and Jichen Yang, “Key-sparse transformer with cascaded cross-attention block for multimodal speech emotion recognition,” arXiv preprint arXiv:2106.11532, 2021.

[14] Liu et al., “Roberta: A robustly optimized bert pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.

[15] Alexei Baevski, Yuahao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” NeurIPS 2020, pp. 12449–12460.

[16] Busso et al., “Iemocap: Interactive emotional dyadic motion capture database,” LREC 2008, pp. 335–359.

[17] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarp Kumar Sarma, and Najim Dehak, “Emotion identification from raw speech signals using dnns.,” in Interspeech 2018, pp. 3097–3101.

[18] Aaron Keesing, Yun Sing Koh, and Michael Witbrock, “Acoustic features and neural representations for categorical emotion recognition from speech,” in Interspeech 2021, pp. 3415–3419.

[19] 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.

[20] Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church, “Speech Emotion Recognition with Multi-Task Learning,” in Interspeech 2021, 2021, pp. 4508–4512.

[21] Diogo Cortiz, “Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra,” arXiv preprint arXiv:2104.02041.

[22] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz J Rak, “Emotion recognition using facial expressions,” Procedia Computer Science, vol. 108, pp. 1175–1184, 2017.

[23] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency, “Context-dependent sentiment analysis in user-generated videos,” in ACL 2017, 2017, pp. 873–883.
[24] Wen Wu, Chao Zhang, and Philip C Woodland, “Emotion recognition by fusing time synchronous and time asynchronous representations,” in IEEE ICASSP 2021, 2021, pp. 6269–6273.

[25] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency, “Words can shift: Dynamically adjusting word representations using nonverbal behaviors,” in AAAI 2019, pp. 7216–7223.

[26] Jilt Sebastian, Piero Pierucci, et al., “Fusion techniques for utterance-level emotion recognition combining speech and transcripts,” in Interspeech 2019, 2019.

[27] Poria et al., “Multimodal sentiment analysis: Addressing key issues and setting up the baselines,” IEEE Intelligent Systems, vol. 33, no. 6, pp. 17–25, 2018.

[28] Woo Yong Choi, Kyu Ye Song, and Chan Woo Lee, “Convolutional attention networks for multimodal emotion recognition from speech and text data,” in 2018 Challenge-HML, 2018, pp. 28–34.

[29] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li, “Learning alignment for multimodal emotion recognition from speech,” arXiv preprint arXiv:1909.05645, 2019.

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

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, “Representation learning with contrastive predictive coding,” arXiv preprint arXiv:1807.03748, 2018.

[32] Vaswani et al., “Attention is all you need,” NeurIPS 2017, vol. 30.

[33] 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 ACL 2019, p. 6558.

[34] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in IEEE ICASSP 2015, pp. 5206–5210.

[35] Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church, “Speech emotion recognition with multi-task learning,” in Interspeech 2021, 2021.

[36] Wu et al., “Speech emotion recognition using capsule networks,” in IEEE ICASSP 2019, pp. 6695–6699.

[37] Muhammad Sajjad, Sooinil Kwon, et al., “Clustering-based speech emotion recognition by incorporating learned features and deep bi lstm,” IEEE Access 2020, vol. 8, pp. 79861–79875.

[38] Zhiyun Lu, Liangliang Cao, Yu Zhang, Chung-Cheng Chiu, and James Fan, “Speech sentiment analysis via pre-trained features from end-to-end asr models,” in IEEE ICASSP 2020, pp. 7149–7153.

[39] Jiaxing Liu, Zhilei Liu, Longbiao Wang, Lili Guo, and Jianwu Dang, “Speech emotion recognition with local-global aware deep representation learning,” in IEEE ICASSP 2020, pp. 7174–7178.

[40] Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh, “Speech emotion recognition with dual-sequence lstm architecture,” in IEEE ICASSP 2020, 2020, pp. 6474–6478.

[41] R. Pappagari, Tianzi Wang, Jesús Villalba, Nanxin Chen, and Najim Dehak, “X-vectors meet emotions: A study on dependencies between emotion and speaker recognition,” IEEE ICASSP 2020, pp. 7169–7173.

[42] Zhichao Peng, Xingfeng Li, Zhi Zhu, Masashi Unoki, Jianwu Dang, and Masato Akagi, “Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends,” IEEE Access 2020, vol. 8, pp. 16560–16572.

[43] Weidong Chen, Xiaofeng Xing, Xiangmin Xu, and Jichen Yang, “Key-sparse transformer with cascaded cross-attention block for multimodal speech emotion recognition,” arXiv preprint arXiv:2106.11532, 2021.

[44] Sarala Padi, Seyed Omid Sadjadi, Dinesh Manocha, and Ram D Sriram, “Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models,” arXiv preprint arXiv:2202.08974, 2022.

[45] Mariana Rodrigues Makiuchi, Kuniaki Uto, and Koichi Shinoda, “Multimodal emotion recognition with high-level speech and text features,” in IEEE ASRU 2021, 2021, pp. 350–357.

[46] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, “Layer-wise analysis of a self-supervised speech representation model,” in 2021 IEEE ASRU, pp. 914–921.

[47] Ghosh et al., “Delores: Decorrelating latent spaces for low-resource audio representation learning,” AAAI 2022 SAS Workshop.

[48] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, “Attentive statistics pooling for deep speaker embedding,” Interspeech 2018.

[49] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli, “Data2vec: A general framework for self-supervised learning in speech, vision and language,” arXiv preprint arXiv:2202.03555, 2022.