INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

MMER: Multimodal Multi-task Learning for Speech Emotion Recognition

Sreyan Ghosh[♠], *Utkarsh Tyagi*[♠], *S Ramaneswaran*[♣], *Harshvardhan Srivastava*[♥], *Dinesh Manocha*[♠]

[♠]University of Maryland, College Park, USA,
[♣]NVIDIA, Bangalore, India, [♥]IIT Delhi, India
{sreyang,utkarsht,dmanocha}@umd.edu

Abstract

In this paper, we propose **MMER**, a novel Multimodal Multi-task learning approach for Speech Emotion Recognition. MMER leverages a novel multimodal network based on early-fusion and cross-modal self-attention between text and acoustic modalities and solves three novel auxiliary tasks for learning emotion recognition from spoken utterances. In practice, MMER outperforms all our baselines and achieves state-of-the-art performance on the IEMOCAP benchmark. Additionally, we conduct extensive ablation studies and results analysis to prove the effectiveness of our proposed approach. [1]

Index Terms: speech emotion recognition, human-computer interaction

1. Introduction

In addition to the explicit messages humans convey, they implicitly express emotions in conversations. Speech Emotion Recognition (SER) aims to identify these implicit emotions from spoken human utterances, which proves to be one of the key components of better human-computer interaction systems.

SER is a well-studied problem in the literature, with a variety of systems proposed that achieve state-of-the-art (SOTA) performance on benchmark datasets [1, 2, 3, 4]. However, most of these systems are uni-modal and learn only from the acoustic modality [5, 6, 7], with very few systems taking a multimodal approach [3, 8]. We emphasize the importance of multimodal learning for SER due to the real-world multi-modal nature of emotional expression in humans, which includes body language, facial expressions, word choice, tone of voice, and more. With Automatic Speech Recognition (ASR) systems achieving near-optimal results, we hypothesize that the modality of text is available as a complementary signal to speech and can significantly improve SER performance by eliminating the natural prosodic bias in spoken utterances.

While work on SER has proposed various learning paradigms that achieve great performance [1], we hypothesize that SER can primarily benefit the most by learning to solve auxiliary tasks that can help infuse extra knowledge into the model. For example, the current SOTA on the IEMOCAP benchmark [9] solves an ASR task with emotion classification, which helps the model learn strong linguistic cues. However, multitask learning is an under-explored area in SER literature, and we emphasize that better auxiliary learning tasks can help the model learn improved representations, thereby improving final SER performance.

Main Contributions. In this paper, we propose MMER, a novel multimodal multitask learning approach for SER. MMER first leverages a *novel multimodal neural network architecture* to capture fine-grained multimodal emotional information from acoustic and text modalities using speech and its corresponding text transcripts. Our proposed architecture captures fine-grained inter-modality interactions and alleviates unimodal biases. Specifically, the model uses strong contextual representations from self-supervised (SSL) models and learns implicit temporal alignments between both modalities using a novel multimodal interaction module, which we discuss in Section 3.2. Next, MMER solves *three auxiliary tasks* in addition to emotion classification. First, it solves an ASR task by minimizing the CTC loss [10] to learn the natural monotonic alignment between speech and text and the semantic and syntactic information hidden in the text. Next, we propose to solve two additional contrastive learning-based tasks: (1) *Supervised Contrastive Learning* (SCL): To enforce the model to learn better emotion features from multi-modal data, we solve a supervised contrastive learning task [11], where we learn instance discrimination with model representations based on ground-truth instance labels. Specifically, for SCL, instances with the same emotion label make up the positives, and those with different labels, make up the negatives. (2) *Augmented Contrastive Learning* (AGL): To make the model more robust to the data and enforce learned features to be speaker invariant, we augment the text using back-translation and generate speech from that text using a speaker-conditioned TTS, conditioned on a different speaker from the training corpus. In practice, MMER achieves SOTA results on the IEMOCAP benchmark for SER. We also perform extensive analysis and ablation studies to prove the effectiveness of each individual component in MMER.

2. Related Work

Unimodal SER. Uni-modal speech-only SER is the most commonly studied system in literature for 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, 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 [12, 13]. With recent advances in self-supervised learning (SSL), pre-trained SSL features, like Natural Language Processing (NLP) [14] 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 [15]. The

[1] Code: https://github.com/Sreyan88/MMER.

10.21437/Interspeech.2023-2271

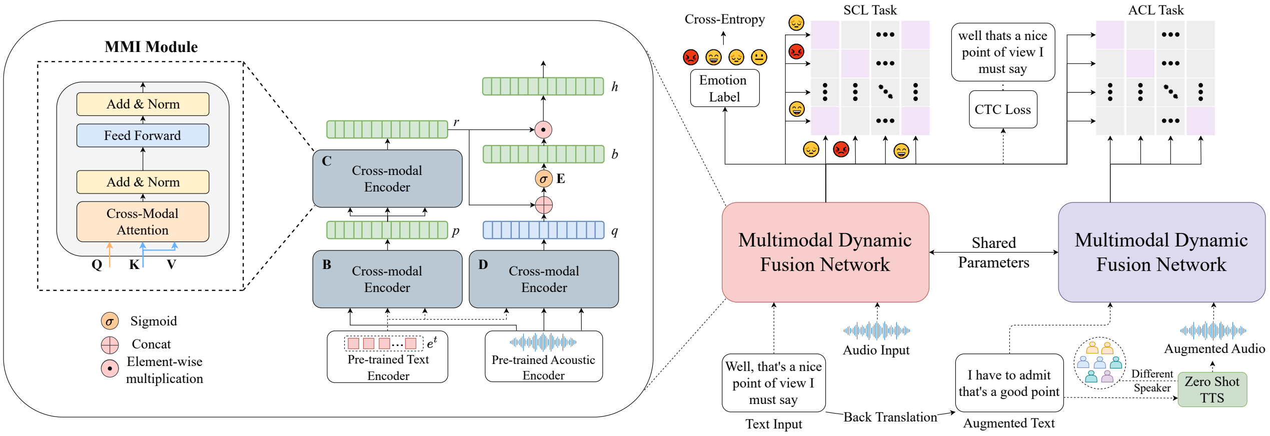

Figure 1: *Illustration of our proposed MMER. MMER introduces a novel Multimodal Dynamic Fusion Network and jointly optimizes 4 different tasks to learn various aspects of SER. We highlight the Multimodal Interaction Module on the left.*

current state-of-the-art on SER [16] also uses wav2vec-2.0 as the speech encoder and solves the SER task 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 the performance of a downstream task when the auxiliary task is chosen properly [17].

Multimodal SER. 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 [18, 19, 20]. Though this technique is simple and effective at modeling modality-specific interactions, it is not effective at modeling cross-modal interactions [21]. Early fusion to capture inter-modality interactions has also been explored [22]. However, in general, early fusion also suppresses modality-specific interactions and does not outperform late fusion methods in emotion recognition [21, 23]. To better model the interactions between modalities, researchers have proposed cross-modal attention (CMA) mechanisms [24, 25, 26, 27]. 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 [24, 25] use dot-product attention, very recently, the use of self-attention-based CMA mechanisms for ER has been gaining traction [26, 28, 27].

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 their 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 belongs to one of the j unique emotions being studied in the dataset.

3.2. Multimodal Dynamic Fusion Network

3.2.1. *Feature Encoder*

Contextualized Speech Representations. To encode speech to obtain high-level contextualized representations, we use a pre-trained wav2vec-2.0 [4] as our raw waveform encoder. 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. For each raw audio input a_i of utterance u_i wav2vec-2.0 outputs $e^{a_i} \in \mathbb{R}^{J\times768}$ where J depends on the length of the raw audio file and the CNN feature extraction layer of wav2vec-2.0, which extracts frames with a stride of 20ms and a hop size of 25ms.

Contextualized Token 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. For a total of M tokens, RoBERTa outputs representations $e^{t_i} \in \mathbb{R}^{M\times768}$. We use RoBERTa only as a feature extractor and do not train it while fine-tuning our model.

3.2.2. *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 is constructed like a generic transformer layer [29], where each layer is composed of an h-head **CMA** module [30], residual connections, and feed-forward layers. In this section, we discuss the working of each of the 3 CME blocks and the acoustic gate E in detail.

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 $\mathbf{A} \in \mathbb{R}^{d\times J}$ as queries and RoBERTa embeddings $\mathbf{T} \in \mathbb{R}^{d\times M}$ as keys and values into the **CMA** module of **CME** block B as follows:

$$\mathbf{CMA}(\mathbf{A},\mathbf{T}) = \mathrm{softmax}\left(\frac{[\mathbf{W_{q_i}A}]^\top [\mathbf{W_{k_i}T}]}{\sqrt{d/m}}\right)[\mathbf{W_{v_i}T}]^\top \tag{1}$$

where $\{\mathbf{W_{q_i}}, \mathbf{W_{k_i}}, \mathbf{W_{v_i}}\} \in \mathbb{R}^{d/m\times h}$ 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 $\mathbf{P} = (\mathbf{p_0},\mathbf{p_1}, \cdots, \mathbf{p_{m-1}})$. Next, to address the fact that each generated representation $\mathbf{p}_i$ in the previous block corresponds to the i^{th} acoustic embedding and not the token embedding, we feed $\mathbf{P}$ to another $\mathbf{CME}$ block C, which treats the original RoBERTa embeddings $\mathbf{T}$ as queries and $\mathbf{P}$ as keys and values. Finally, we obtain the final Speech-Aware Word Representations as $\mathbb{R} = (\mathbf{r_0},\mathbf{r_1}, \cdots, \mathbf{r_{j-1}})$.

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 $\mathbf{CME}$ block D by treating $\mathbf{T}$ as queries and $\mathbf{A}$ as keys and values. The final representations obtained from the block can be denoted as $\mathbf{Q} = (\mathbf{q_0},\mathbf{q_1}, \cdots, \mathbf{q_{j-1}})$. 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.

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 E that can dynamically control the contribution of each speech frame embedding. Following previous work, we implement an acoustic gate $\mathbf{g}$ as follows:

$$\mathbf{g} = \sigma\left(\mathbf{W}_g^\top[\mathbf{R};\mathbf{Q}] + \mathbf{B}_g\right) \tag{2}$$

where $\mathbf{W}_g \in \mathbb{R}^{2dXd}$ is a weight matrix, $\mathbf{B}_g \in \mathbb{R}^d$ is the bias, and σ is the element-wise sigmoid function. Finally, based on the gate output, the final word-aware speech representations are obtained by $\mathbf{Q} = \mathbf{g}.\mathbf{Q}$. After this step, we concatenate the speech-aware word representations and word-aware speech representations to obtain our final cross-modal MMI representations $\mathbf{M} \in \mathbb{R}^{2d}$ where $\mathbf{M} = [\mathbf{Q} ; \mathbf{R}]$, and pass it through a linear transformation $l(.)$, which down-projects $\mathbf{M}$ to a d dimensional space.

3.3. Multi-task Learning

As discussed earlier, MMER solves a total of 4 tasks for learning SER. In this sub-section, we briefly discuss all these losses and their contribution to learning SER.

Cross-Entropy Loss. Cross-Entropy is the most common loss used for learning SER. To calculate the Cross-Entropy Loss with ground-truth emotion labels, we first employ max pooling $\mathbf{mp}(.)$ over wav2vec-2.0 speech representations ($\mathbf{A}$) and MMI module ($\mathbf{M}$) 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:

$$\hat{y} = \mathbf{softmax}\left(\mathbf{W}_p^\top[\mathbf{mp}(\mathbf{A}); \mathbf{mp}(\mathbf{M})] + \mathbf{B}_p\right) \tag{3}$$

where $\hat{y} \in \mathbb{R}^4$ is the single vector representation for each utterance, $\mathbf{W}_p \in \mathbb{R}^{2d\times 4}$ is a weight matrix, $\mathbf{softmax}(.)$ denotes the softmax activation function, and $\mathbf{mp}(.)$ denotes the attention pooling operation across the embedding axis. After this step, Cross Entropy is calculated by $\mathcal{L}_{\text{CE}} = \text{CrossEntropy}(\hat{y_i}, y_i)$.

CTC Loss. MMER next solves the ASR task by minimizing the CTC loss. Learning to solve the ASR task encourages the model to learn linguistic properties like speech and text by leveraging the natural monotonic alignment between the acoustic and textual modalities. To do this, we first pass the raw un-pooled embeddings $\mathbf{A}$ from the wav2vec-2.0 encoder through a linear layer as follows:

Algorithm 1 Supervised & Augmented Contrastive Learning

```
Require: A list of emotion labels for all data in a batch is L; each emotion
    is divided into four categories; The Multimodal Dynamic Fusion Network is
    MDFN; the texts are T; the audio files are A; BT denotes back-translation,
    and TTS denotes zero-shot speaker-conditioned text-to-speech; $\mathrm{E}_{\text{speaker}}$ are
    the speaker embeddings for speakers in batch; $C$ denotes the length of $L_c$;
    $S$ denotes the length of $L$; $\mathcal{L}_{\text{SCL}}$ denotes Supervised contrastive loss $\mathcal{L}_{\text{ACL}}$
    denotes augmented contrastive loss.
    Initialize $L_c = [L - 0, \ldots, L - 4]$ and $L_t = list()$
    for $i = 1; i <= C; i + +$ do
        initialize $\tilde{L}_t = list()$
        for $j = 1; j <= T; j + +$ do
            if $L_c[i][j]$ equals 0 then
                $\tilde{L}_t.append(j)$
            end if
        end for
        $L_t.append(\tilde{L}_t)$
    end for
    $R = \text{MDFN}(T, A)$
    $\tilde{T} = \text{BT}(T)$
    $R_{au} = \text{MDFN}(\tilde{\text{T}}, \text{TTS}(\tilde{T}, \mathrm{E}_{\text{speaker}}))$
    $\tilde{l}_{pn} = \text{einsum}(nc, ck \rightarrow nk, [R, RT])$
    $l_{pn} = \text{LogSoftmax}(\tilde{l}_{pn}/\tau))$
    $L_{cl} = L_t[L[1]]$
    for $q = 2; q <= S, q + +$ do
        $L_{cl} = concat(L_{cl}, L_t[L[q]] + qT)$
    end for
    $\mathcal{L}_{\text{SCL}} = \text{gather}(l_{pn}, index = L_{cl})/T$
    $l_{pn} = \text{einsum}(nc, ck \rightarrow nk, [R, RT_{au}])$
    $cl_{label} = \text{arange}(S)$
    $\mathcal{L}_{\text{ACL}} = \text{CrossEntropy}(l_{pn}/\tau, cl_{label})$
    return $\mathcal{L}_{\text{SCL}}$ $\mathcal{L}_{\text{ACL}}$
```

$$\hat{t} = \mathbf{softmax}\left(\mathbf{W}_c^\top \mathbf{A} + \mathbf{B}_c\right) \tag{4}$$

where $\hat{t} \in \mathbb{R}^{J\times V}$, 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. $\mathbf{W}_c \in \mathbb{R}^{d\times V}$ and $\mathbf{B}_c$ is the added bias. After this step; we calculate the CTC loss by $\mathcal{L}_{\text{CTC}} = \text{CTC}(\hat{t}, t)$, where $t_i \in \{t_0,\cdots,t_i,\cdots,t_N\}$ is a pre-processed version of the original t_i where we remove all punctuation and convert all characters to uppercase.

Supervised Contrastive Learning. Supervised Contrastive Learning (SCL) [11] supplements the Cross-Entropy to learn better emotion features. Precisely, SCL solves the generic instance discrimination contrastive learning task with multimodal representations $\mathbf{M}$, but in the presence of emotion labels. To solve SCL, we first divide the representations in each batch into multiple subsets according to their emotion label. Then, for each subset, representations within that subset act as the positives, while representations in another subset act as the negatives. Fig. 1 illustrates the process, and we show specific steps in Algorithm 1.

Augmented Contrastive Learning. Augmented Contrastive Learning (ACL) encourages MMER to learn invariant features in the data. Past work has shown that SER benefits from learning speaker invariance [31]. However, in this work, we take a slightly different approach to consider the multimodal nature of MMER and additionally learn semantic invariance in text. Thus, MMER solves another instance discrimination contrastive learning task between multimodal representations $\mathbf{M}$, where the representations are learned from the augmented text and speech cues. To augment the text, we use back-translation, which refers to translating an existing text into a target language and then back into the source language. Yu *et al.* [32] show that back-translation can generate diverse sentences while preserving the semantics of the original sentence. For speech, our primary objective is to generate an augmented utterance of the same emotion but as if uttered by a different speaker. To perform this, we use a SOTA zero-shot speaker-conditioned TTS [33] and generate utterances from the back-translated text. The system proposed by Casanova *et al.* [33] takes speaker embeddings as input added to the text, and we calculate these embeddings with all utterances from a speaker randomly sampled from the dataset but expressing a similar emotion. Fig. 1 illustrates the process, and we show specific steps in Algorithm 1. Finally, for optimizing MMER we minimize $\mathcal{L}$ as: $\mathcal{L} = \mathcal{L} + \alpha\mathcal{L}_{\text{CTC}} + \beta\mathcal{L}_{\text{SCL}} + \gamma\mathcal{L}_{\text{ACL}}$ where α, β and γ are hyper-parameters that we tune.

4. Experiments and Results

Dataset. Following much of the prior work in SER literature, we train and evaluate all our models on the IEMOCAP dataset [9]. IEMOCAP contains approximately 12 hours of speech from a total of 10 speakers, all of which come from 5 scripted sessions, acted by professional actors. To keep our dataset settings consistent with the prior work 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 labeled 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.

Baselines. We build unimodal baselines with just text and speech modalities, where the text baseline uses RoBERTa$_{\text{BASE}}$ as the contextualized text encoder, followed by a single linear layer and softmax activation for classification. For the unimodal speech baseline, we use the same setup but replace our encoder with pre-trained wav2vec-2.0-base, pre-trained on 960hrs of LibriSpeech [34]. We also build a naive multimodal baseline where we simply concatenate pooled self-supervised representations $\mathbf{mp}(\mathbf{A})$ and $\mathbf{mp}(\mathbf{T})$ in a single-task SER learning setup. We compare our model with other methods in the literature evaluated on 5-fold cross-validation setups, including unimodal and multimodal approaches. All results for prior art have been taken from the literature (weighted accuracy unless stated otherwise). We only re-implement the current state-of-the-art approach [1] under the 5-fold cross-validation setup for a fair comparison.

Hyper-parameters. 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 4 and accum-grad of 4 for 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 $\alpha, \beta, \gamma = 0.1$ where the search was performed among $\alpha, \beta, \gamma \in \{1, 0.1, 0.01, 0.001\}$ with grid search. Each training and inference step took 10 minutes on a single NVIDIA A100 GPU. MMER has ≈228M parameters.

Quantitative Analysis. Table 1 compares the performance of MMER with all our baselines. MMER achieves SOTA performance on the IEMOCAP benchmark, with the closest being [2], where the author uses 2 contextualized speech encoders, resulting in more than double the number of parameters as ours. MMER also benefits from minimal trainable parameter addition over [1] or a simple wav2vec-2.0. We achieved 75.0% WA when Google transcripts were used instead of gold transcripts for inference.

Qualitative Analysis. Fig. 2 shows the confusion matrix for various settings with or without a particular objective function. One clear observation is that ACL alleviates the bias to the neutral class, which is a common problem in prior art [35]. On the contrary, CTC amplifies this by a small amount, which we attribute to the fact that our model learns more semantic information in text, ignoring important cues in speech. We provide results on various settings of α, β, and γ on our GitHub.

Table 1: *Emotion Recognition Results on IEMOCAP*

Method	CV	Modality	WA
Prior-art			
Wu et al. [5]	10-fold	{a}	72.7%
Sajjad et al. [6]	5-fold	{a}	72.3%
Lu et al. [7]	10-fold	{a}	72.6%
Liu et al. [36]	5-fold	{a}	70.8%
Wang et al. [37]	5-fold	{a}	73.3%
Zhao et al. [38]	5-fold	{a,t}	76.3%
Yang et al. [39]	5-fold	{a,t}	77.7%
Morais et al. [2]	5-fold	{a,t}	77.4%
Chen et al. [35]	5-fold	{a,t}	74.3%
Padi et al. [3]	5-fold	{a,t}	75.0%
Makiuchi et al. [8]	5-fold	{a,t}	73.5%
Chen et al. [27]	5-fold	{a,t}	74.3%
Cai et al. [1]	10-fold	{a,t}	77.1%
Our Baselines			
RoBERTa$_{BASE}$	5-fold	{t}	69.2%
wav2vec-2.0	5-fold	{a}	73.9%
Multimodal	5-fold	{a,t}	74.1%
Proposed			
MMER w/o CTC	5-fold	{a,t}	78.1%
MMER w/o SCL	5-fold	{a,t}	78.9%
MMER w/o ACL	5-fold	{a,t}	79.8%
MMER	**5-fold**	**{a,t}**	**81.2%**

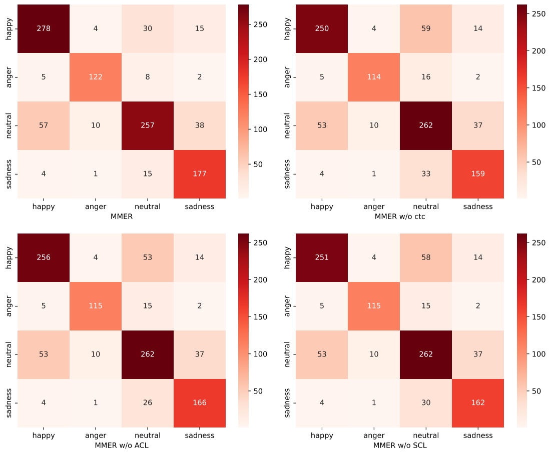

Figure 2: *Confusion Matrix: Ablation study and MMER performance evaluation with and without the CTC loss and the proposed objective functions, SCL and ACL..*

5. Conclusion and Limitations

In this paper, we propose MMER, a novel multimodal multitask approach for SER from spoken utterances. MMER leverages a novel dynamic multimodal fusion network and three additional auxiliary tasks. As part of future work, we would like to work on the current limitations of MMER, including the requirement of large batch sizes for training in contrastive learning and pre-computed text features.

6. References

[1] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in *Interspeech 2021*, 2021.

[2] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," *arXiv preprint arXiv:2202.03896*, 2022.

[3] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models," *arXiv preprint arXiv:2202.08974*, 2022.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS 2020*, pp. 12 449–12 460.

[5] W. et al., "Speech emotion recognition using capsule networks," in *IEEE ICASSP 2019*, pp. 6695–6699.

[6] M. Sajjad, S. Kwon *et al.*, "Clustering-based speech emotion recognition by incorporating learned features and deep bilstm," *IEEE Access 2020*, vol. 8, pp. 79 861–79 875.

[7] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, "Speech sentiment analysis via pre-trained features from end-to-end asr models," in *IEEE ICASSP 2020*, pp. 7149–7153.

[8] M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal emotion recognition with high-level speech and text features," in *IEEE ASRU 2021*, 2021, pp. 350–357.

[9] B. et al., "Iemocap: Interactive emotional dyadic motion capture database," *LREC 2008*, pp. 335–359.

[10] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[11] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[12] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using dnns." in *Interspeech 2018*, pp. 3097–3101.

[13] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic features and neural representations for categorical emotion recognition from speech," in *Interspeech 2021*, pp. 3415–3419.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] Y. et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech 2021*, 2021, pp. 1194–1198.

[16] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech Emotion Recognition with Multi-Task Learning," in *Interspeech 2021*, 2021, pp. 4508–4512.

[17] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," *arXiv preprint arXiv:1804.10816*, 2018.

[18] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," *arXiv preprint arXiv:1804.05788*, 2018.

[19] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *ACL 2017*, 2017, pp. 873–883.

[20] W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *IEEE ICASSP 2021*, 2021, pp. 6269–6273.

[21] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *AAAI 2019*, pp. 7216–7223.

[22] J. Sebastian, P. Pierucci *et al.*, "Fusion techniques for utterance-level emotion recognition combining speech and transcripts." in *Interspeech 2019*, 2019.

[23] P. et al., "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[24] W. Y. Choi, K. Y. Song, and C. W. Lee, "Convolutional attention networks for multimodal emotion recognition from speech and text data," in *2018 Challenge-HML*, 2018, pp. 28–34.

[25] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," *arXiv preprint arXiv:1909.05645*, 2019.

[26] D. Krishna and A. Patil, "Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks." in *Interspeech 2020*, 2020, pp. 4243–4247.

[27] W. Chen, X. Xing, X. Xu, and J. Yang, "Key-sparse transformer with cascaded cross-attention block for multimodal speech emotion recognition," *arXiv preprint arXiv:2106.11532*, 2021.

[28] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," *arXiv preprint arXiv:2009.04107*, 2020.

[29] V. et al., "Attention is all you need," *NeurIPS 2017*, vol. 30.

[30] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *ACL 2019*, p. 6558.

[31] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *IEEE ICASSP 2020*, pp. 7144–7148.

[32] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "Qanet: Combining local convolution with global self-attention for reading comprehension," in *ICLR 2018*.

[33] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *ICML 2022*, pp. 2709–2720.

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE ICASSP 2015*, pp. 5206–5210.

[35] W. Chen, X. Xing, X. Xu, and J. Yang, "Key-sparse transformer with cascaded cross-attention block for multimodal speech emotion recognition," *arXiv preprint arXiv:2106.11532*, 2021.

[36] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *IEEE ICASSP 2020*, 2020, pp. 7174–7178.

[37] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence lstm architecture," in *IEEE ICASSP 2020*, 2020, pp. 6474–6478.

[38] Z. Zhao, Y. Wang, and Y. Wang, "Multi-level fusion of wav2vec 2.0 and bert for multimodal emotion recognition," 2022. [Online]. Available: https://arxiv.org/abs/2207.04697

[39] W. Yang, S. Fukayama, P. Heracleous, and J. Ogata, "Exploiting Fine-tuning of Self-supervised Learning Models for Improving Bi-modal Sentiment Analysis and Emotion Recognition," in *Proc. Interspeech 2022*, 2022, pp. 1998–2002.