scientific reports

Check for updates

OPEN

MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion

Mustaqeem Khan[1,6], Phuong-Nam Tran[2,6], Nhat Truong Pham[3,6], Abdulmotaleb El Saddik[1,4] & Alice Othmani[5]✉

Speech emotion recognition has seen a surge in transformer models, which excel at understanding the overall message by analyzing long-term patterns in speech. However, these models come at a computational cost. In contrast, convolutional neural networks are faster but struggle with capturing these long-range relationships. Our proposed system, *MemoCMT*, tackles this challenge using a novel "cross-modal transformer" (*CMT*). This *CMT* can effectively analyze local and global speech features and their corresponding text. To boost efficiency, *MemoCMT* leverages recent advancements in pre-trained models: *HuBERT* extracts meaningful features from the audio, while *BERT* analyzes the text. The core innovation lies in how the *CMT* component utilizes and integrates these audio and text features. After this integration, different fusion techniques are applied before final emotion classification. Experiments show that *MemoCMT* achieves impressive performance, with the *CMT* using min aggregation achieving the highest unweighted accuracy (*UW-Acc*) of 81.33% and 91.93%, and weighted accuracy (*W-Acc*) of 81.85% and 91.84% respectively on benchmark IEMOCAP and ESD corpora. The results of our system demonstrate the generalization capacity and robustness for real-world industrial applications. Moreover, the implementation details of *MemoCMT* are publicly available at https://github.com/tpnam0901/MemoCMT/ for reproducibility purposes.

Keywords Multimodal emotion recognition, Speech emotion recognition, Cross-modal transformer, Deep learning, Feature fusion

Multimodal emotion recognition (*MER*) goes beyond traditional emotion recognition, which looks solely at the content of spoken words. This technology analyzes what we say and how we say it. By incorporating tone of voice, facial expressions, and even text transcript, *MER* paints a richer picture of our emotions during communication. This holds immense potential for various fields like healthcare, education, and customer service, paving the way for more nuanced and empathetic interactions between humans and machines[1]. As highlighted in[2], there's a growing need for advanced systems that process speech, video, and text to recognize emotions. Human speech is rich with emotional cues, allowing us to effortlessly convey our feelings during communication. This significant advancement has led to the creation of speech emotion recognition (*SER*) systems, representing a crucial step forward in understanding and interpreting human emotions through speech, which hold immense potential for real-world applications across various fields[3]. These applications include robotics, security, language translation, automated identification systems, intelligent toys, and even lie detection[1].

Despite significant progress in speech processing, achieving high accuracy, real-time emotion recognition for practical use remains a challenge[4,5]. Researchers are leveraging cutting-edge technologies like attention-based approaches[6,7] to improve speaker identification by analyzing their emotional state during speech. In contrast, researchers and scientists implemented several systems for emotion recognition via joint features learning-based methods[8–10] to enhance the recognition rate and tried text modality[3,11] to further increase the model performance for real-time applications. Hence, Siriwardhana[12] used a transformer-based pre-trained model to combine speech and text modalities, which are essential for recognizing emotions and improving the precision rate. Similarly, Feng et al.[8] and Chen et al.[13] integrated a speech emotion with automatic speech recognition (*ASR*) to make the system more intelligent for real-world applications. However, optimizing pre-trained models

[1]College of Information Technology, United Arab Emirates University (UAEU), Al Ain, Abu Dhabi, United Arab Emirates, 5551 Al Ain, UAE. [2]Department of Artificial Intelligence, Kyung Hee University, Yongin-si 17104, Republic of Korea. [3]Sungkyunkwan University, Suwon, Gyeonggi-do 16419, Republic of Korea. [4]University of Ottawa, Ottawa K1H 8M5, Canada. [5]LISSI Laboratory, University Paris-Est Créteil, 94400 Vitry sur Seine, France. [6]Mustaqeem Khan, Phuong-Nam Tran and Nhat Truong Pham contributed equally to this work. ✉email: alice.othmani@u-pec.fr

for *SER* is complex, and further research and exploration are necessary[14] to reduce the latency with a high recognition rate. A recent trend of deep learning (DL)-based attention mechanism rapidly growing in speech processing to improve performance by focusing on salient cues in speech and text sequence[15–17].

In this regard, Xu et al.[18] and Naderi and Nasersharif[19] introduced a multi-head self-attention module to improve recognition rate, and[20] used local attention to learn emotions automatically with high precision by a similar technique. The above-discussed approaches are black-box[21], and their internal mechanisms are impossible to understand. Recent research has focused on explaining the internal workings and mechanism of DL-based techniques[22] to understand multimodal emotion using biological signals according to Lin et al.[23]. The work in[24] developed a model to evaluate different emotional elements and interpret their response for specific tasks. Hence, a method was developed by[25] to identify the component input tensor responsible for a specific output. Therefore, Shrikumar et al.[26] developed a technique for decomposing output predictions by tracking the contributions of each neuron. The network training must be conducted layer by layer, as this approach is not visible in current methods. This insight provides a solid foundation for creating an innovative DL technique to accurately interpret and decipher predictions for the speech emotion system. Adopting this layered training approach can enhance the model's ability to capture hierarchical features in speech signals, potentially leading to more nuanced emotion recognition.

This method allows for fine-tuning individual layers, ensuring that each level of abstraction is optimized for emotion-relevant features. Furthermore, it enables better transparency and interpretability of the model's decision-making process, addressing the 'black box' problem often associated with deep learning systems. The layer-by-layer training also facilitates the integration of domain-specific knowledge at different stages of the network, potentially improving the overall performance and generalization of the *SER*. Additionally, this technique may lead to more efficient training procedures, allowing for targeted optimization of specific layers without retraining the entire network. This could result in reduced computational requirements and faster model iterations. The insights gained from this layered training method could also be extended to other multimodal emotion recognition tasks, paving the way for more sophisticated and accurate affective computing systems.

In this study, we propose *MemoCMT*, a cross-modal transformer-based fusion method for *MER*, which leverages both audio and textual feature representations through a cross-modal transformer (*CMT*) mechanism (Fig. 1). Notably, we utilize the *HuBERT*[27] network in the *SER* module and the *BERT*[28] network in the text emotion recognition (*TER*) module. More specifically, the acoustic characteristics of the speech have been extracted via a pre-trained *HuBERT* model, and textual cues are extracted through a pre-trained *BERT* model. Moreover, investigate different *MemoCMT* design configurations to achieve optimal contextual modeling using a *CMT* inspired by cross-attention module[19] and various aggregation techniques (including class token (CLS), mean aggregation (MEAN), min aggregation (MIN), and max aggregation (MAX)). It should be noted that the pre-trained *HuBERT* model can simultaneously learn and align audio with its transcripts during pre-training. By leveraging feature fusion via the *CMT* mechanism, *MemoCMT* captures both phonetic and linguistic features from the audio and aligns this information with the contextualized text representations from *BERT*. This results in meaningful and robust feature representations for *MER*. Our analysis discovered that *CMT* with MIN is the

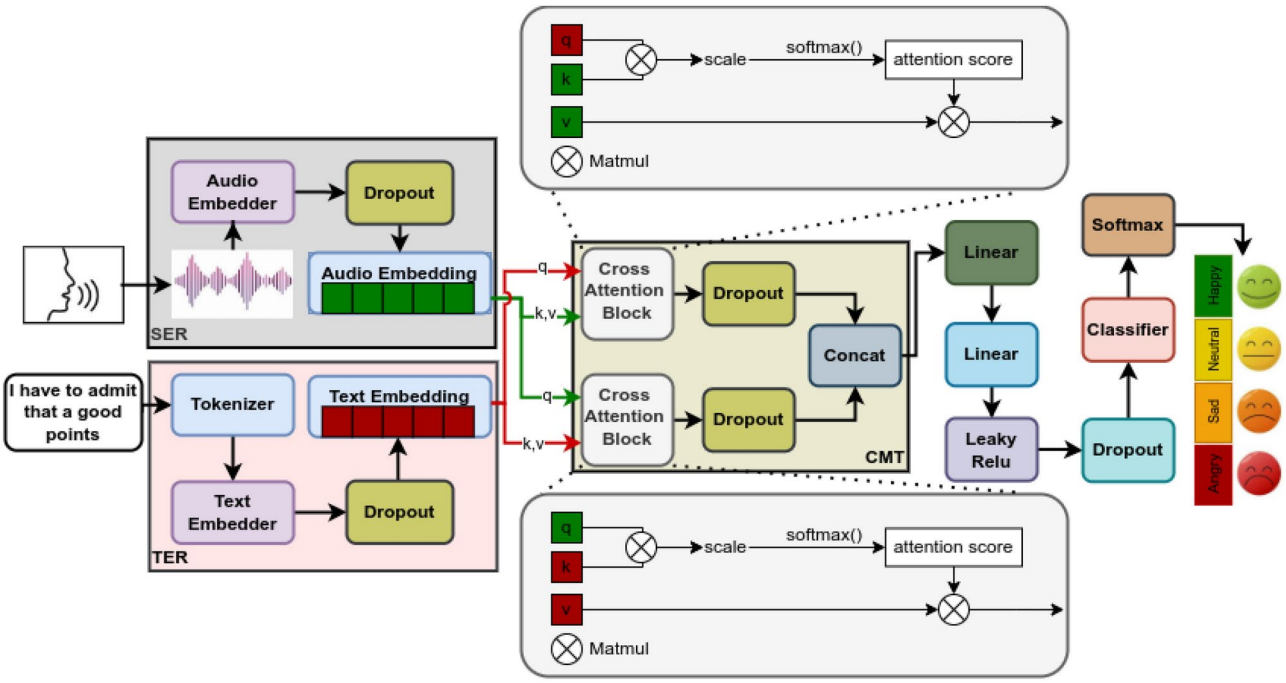


Fig. 1. Overview of the proposed (*MemoCMT*) architecture with the updated *CMT* that leverages cross-attention mechanism to effectively fuse audio and textual representations extracted by the *SER* and *TER* modules, respectively, with a focus on emotional cues.

most efficient design to perform better than other sequential designs (See Section: "Experimental setup and results"). Furthermore, emotion embedding plots have been utilized to visually represent the overlap of various emotion categories based on the t-SNE[29], providing a quantitative analysis of model training (See Section: "Experimental setup and results"). The key points of our *MemoCMT* are the following:

- We propose a novel fusion module called *CMT*. Using a cross-attention module, this module aims to extract key emotional features from audio and text cues. The *CMT* creates a better insight feature that enables the classifier to have more important information for the predicted class.
- To further improve the extracted emotional features, we experiment with various aggregation techniques before passing the fusion feature to the classifier. This method helps choose the suitable reduction feature dimensions created by a cross-attention mechanism for the classifier.
- Our system demonstrates exceptional performance, achieving baseline results on the test sets of three widely recognized benchmark datasets: IEMOCAP, ESD, and MELD. This superior performance is consistently observed across all evaluation metrics, underscoring the robustness and reliability of our approach.The rest of the article is organized as follows: The second section represents the recent literature about multimodal *SER*, and the third section illustrates the main framework and related text and audio encoder. The fourth section reports the corpora information and qualitative and quantitative results of the system with model configuration. Finally, the article concludes in the fifth section with possible future directions.

Recent literature

Multimodal approaches have shown great promise in *SER* by incorporating information from various sources like speech, video, and text. However, a significant challenge remains: effectively combining these modalities[5]. As pointed out in[30], there are key differences between modalities. Some modalities might be more independent than others, and the information they convey might be synchronized or asynchronous. This makes seamless integration a critical research area in multimodal *SER*. Traditionally, researchers have focused on identifying the optimal stage within the model architecture to combine features from different modalities[30].

The four primary types of fusion techniques[31] are fusions based on components/features (often called early fusions), fusions based on decisions (late fusions), fusions based on models, and hybrid fusions. Feature-level fusion encompasses amalgamating features from diverse modalities including visual, text, and audio, consolidating them into comprehensive feature vectors. These vectors are subsequently leveraged for analytical purposes. This approach leverages low-level data features early, including a more extensive range of information from the original data[32]. Researchers have developed the Parallel Inception Convolution Neural Network to improve deep learning methods in pursuit of better methods. Using the concatenation method, features of varying scales are combined to form a standard convolution neural network by simultaneously processing sigma signals from six channels[33]. Numerous researchers have effectively integrated audio, video, and text features by condensing them and channeling them into a Transformer Encoder[34]. Nevertheless, feature-level fusion comes with certain limitations. Various modalities are often represented in the features obtained using this technique. They may exhibit disparities in numerous aspects, necessitating the conversion of these features into a uniform format before fusion[35]. A high-dimensional feature set may also suffer from data sparsity issues because this fusion method lacks interaction of intra-modality information[36]. This can lead to redundancy in modal information and potential overfitting of the data. In summary, while feature-level fusion offers advantages, it is not exempt from limitations.

To overcome the limitations associated with feature-level fusion, decision-level fusion integrates unimodal decision values through ensemble learning techniques such as tensor fusion[17,37], or multiplication layer fusion[31]. Each modality's features are individually analyzed and classified in this approach, and the resulting decisions are combined into decision vectors to yield the final output. Decision-level fusion offers several advantages, simplifying the decision-making process across different modalities compared to feature-level fusion, as multiple modalities often share the same data format[31]. Also, it lets each modality use its best classifier or model for learning features[38]. However, utilizing distinct classifiers or models in the analysis task introduces complexity and increases the time requirements for the learning process during the decision-level fusion phase. Additionally, this approach must address the challenge of capturing the subtleties of modal dynamics without considering the interaction and correlation between different modalities.

Unlike many existing methods that rely on a single fusion strategy or model, our approach leverages a multi-level fusion strategy to capture the intricate interplay of information within and between different modalities (speech and text, in this case). These results are superior to those of baseline methods. Our system focuses on recognizing emotions primarily from speech and text data. The novelty lies in combining both feature-level and model-level fusion techniques. In addition, we are introducing a groundbreaking Model-Fusion module meticulously crafted to enable seamless interactions between modalities and within each modality itself. This allows us to employ a unified model that captures the subtle and dynamic relationships between speech and text features, ultimately leading to more accurate emotion recognition.

Development of MemoCMT

The architecture of the designed *MER* system is depicted in Fig. 1, seamlessly integrating both audio and text modalities to successfully discern emotions. Data pre-processing, feature extraction, modalities fusion, and emotion classification are involved in this section. The effectiveness of *MER* depends on factors such as data quality, feature selection, and the classification method utilized to ensure its resilience. In the leading architecture, we have *SER* and *TER* modules to extract the corresponding cues from each modality. The proposed system is designed to effectively capture and fuse audio and textual cues. Initially, a significant amount of audio and text data is employed from the public benchmark datasets and pre-processed for the relevant modules. The audio is

resampled at a sample rate of 16kHz, and the text is applied to a regular expression to remove redundant space and special characters. Following data pre-processing, the methodology adopts a modality-specific approach, using separate neural networks for audio and text data. These networks learn to extract and represent modality-specific features.

The modality-specific networks, *BERT* and *HuBERT*, are jointly trained via a *CMT* in the next stage. *HuBERT* was initially derived from the *BERT* architecture, which focuses on extracting features from audio data. *BERT* is commonly utilized to extract textual features. The similarity between these architectures can potentially enhance feature representation for both audio and text, facilitating a more seamless alignment in the cross-attention module during subsequent stages. This allows the model to learn the complementary information from audio and text modalities. The training process is fine-tuned to ensure optimal model performance. Cross-validation techniques rigorously evaluate the model's performance, considering weighted accuracy (*W-Acc*) and unweighted accuracy (*UW-Acc*) metrics. Further consideration is given to explainability to ensure the model can be interpreted and provides insights into its predictions. This can involve techniques like aggregation and contextualization mechanisms or visualization of the contributions of different input parts to the final prediction. Ultimately, the *MER* system improves its accuracy and robustness by using *BERT* text-based features and *HuBERT* audio-based features. Further detailed explanations of each module are illustrated in the subsequent sections.

Module 1 (SER)

During this phase, the system inputs an audio waveform to extract important acoustic features. To effectively analyze speech data for emotions, we utilize *HuBERT*[27], which represents audio features using a self-attention mechanism and masked prediction of hidden units technique. The original *HuBERT* architecture has been frozen during the training phase to keep the useful features learned from the LibriSpeech[39] corpus. The LibriSpeech corpus consists of 1000 hours of speech sampled at 16kHz, sourced from various audiobooks. This makes *HuBERT* well-suited for capturing the subtle emotional nuances within the audio signal.

To begin with, *HuBERT* breaks down the sound waves into smaller segments known as tokens using a *CNN* encoder. This step generates numerous tokens, each with a feature-length of 768. By representing the audio waveform as tokens, *HuBERT* can now utilize a transformer architecture to understand the audio features. The transformer architecture of *HuBERT* is built upon the *BERT* architecture but with an extra step of masking tokens before feeding them into *BERT*. The strategic utilization of masked tokens in *HuBERT* undeniably plays a crucial and indispensable role in the overall efficacy of the training process. These tokens are randomly selected and hidden from the model during training. Using masked tokens allows us to leverage the surrounding context more effectively in *HuBERT* learns to understand the audio better and can generalize its knowledge to unseen data.

Module 2 (TER)

The training of the *BERT* model, as described in[28], involves providing the model with the text transcriptions of the spoken utterances, which are initially utilized in the *TER* phase. *BERT* employs a 12-transformer block configuration with 12 attention heads and a pre-trained model featuring 768 hidden units. This configuration is employed to generate encoded hidden vectors r_i for each token represented by the e_i embedding, where *i* refers to the index of the input token. Furthermore, unique tokens c_1 and c_2 are introduced to serve as identifiers for a classifier and separator, which play a role in a specific task.

For each data point represented by *i*, with its corresponding time steps denoted as *t*, the input tokens are obtained through the *BERT* module, resulting in a text feature vector. This task-specific *BERT* exhibits its remarkable capability to provide deep bidirectional representations by randomly masking specific input tokens and then predicting these masked tokens based on the context of the remaining tokens. This methodology allows *BERT* to acquire a profound understanding of the contextual relationships between words within a sentence, thereby enabling it to capture the semantic meaning of the text more accurately than traditional language models and methods. This *BERT* model must be trained to enhance the system's ability to recognize speech emotions.

Module 3 (CMT)

CMT is a powerful module that leverages the cross-attention architecture[19] to fuse speech and text cues effectively. Figure 1 illustrates the architecture of *CMT*. In *CMT*, cross-attention is achieved through a mechanism called multi-head attention[36]. This mechanism takes a set of query-key pairs and a value as input and produces an output, where the *query, key, value,* and output are all vectors. Let denotes $v_s \in R^{n \times d}$ and $v_t \in R^{m \times d}$ as the feature vectors created by *SER* and *TER*, respectively, where *n* indicate the speech cues, *m* indicate the text cues and *d* is the feature dimensional of output vector, Q_s, K_s, V_s indicated the query, key, and value of speech embedding, while Q_t, K_t, V_t correspond to the query, key, and value of text embedding. The *CMT* formula can be defined as follows:

$$CMT_{output} = Concat(Cross\ Attention(Q_s, K_t, V_t), Cross\ Attention(Q_t, K_s, V_s)) \quad (1)$$

where $Q_s = K_s = V_s = v_s$ and $Q_t = K_t = V_t = v_t$. The cross-attention module[19] in this equation helps enable *CMT* seamless integration and extraction of information from speech and text data sources. The formula of cross-attention is as follows:

$$Cross\ Attention^{i\times d} = LayerNorm\left[softmax\left(\frac{(Q^{i\times d}\cdot W_q^{d\times d})\cdot(K^{j\times d}\cdot W_k^{d\times d})^T}{\sqrt{d}}\right)\cdot(V^{j\times d}\cdot W_v^{d\times d})\right] \cdot W^{d\times d} + b^d \quad (2)$$

where *i*, *j* are the number of vector features, W_q, W_k, W_v, W, b are trainable parameters. From Eqs. (1) and (2), the final output of *CMT* is as follow:

$$CMT_{output}^{(m+n)\times d} = Concat\left\{LayerNorm\left[softmax\left(\frac{(Q_s^{n\times d}\cdot W_q^{d\times d})\cdot(K_t^{m\times d}\cdot W_k^{d\times d})^T}{\sqrt{d}}\right)\cdot(V_t^{m\times d}\cdot W_v^{d\times d})\right]\cdot W^{d\times d} + b^d,\right.$$
$$\left.LayerNorm\left[softmax\left(\frac{(Q_t^{m\times d}\cdot W_q^{d\times d})\cdot(K_s^{n\times d}\cdot W_k^{d\times d})^T}{\sqrt{d}}\right)\cdot(V_s^{n\times d}\cdot W_v^{d\times d})\right]\cdot W^{d\times d} + b^d\right\} \quad (3)$$

where LayerNorm[40] is a technique to normalize the vector based on its values. The formula does not include the multi-head process for simplification, where the cross-attention is split into multiple smaller heads to compute. In reality, the multi-head process enhances the model's ability to capture vario0us aspects of the input data. The outputs of these heads are combined to create the final output of the cross-attention mechanism.

The *CMT* generates attention fusion feature vectors $v_{cmt} \in R^{(m+n)\times d}$ where $(m+n)$ is the number of vectors. The v_{cmt} represents the information of speech and text, which was captured using *HuBERT* and *BERT*, respectively. Before passing this v_{cmt} to the classifier, we need to reduce the number of vectors in v_{cmt} into one unique vector. A simple method is to flatten the v_{cmt} to create a unique vector. However, this process creates many units in the linear layer, leading to high computation and bottlenecks. To avoid this issue, we experiment with four aggregation methods to reduce the v_{cmt} vectors: *CLS* method[28], MEAN, MAX, and MIN. CLS selects the first token as a compact summary for classification. MEAN calculates the average value along the token axis, reducing the feature vector dimensions. MAX determines the highest value, while MIN captures the lowest value along the token axis. These aggregation methods produce a smaller $v_{cmt} \in R^d$ compared to the flattened fusion feature, addressing the computational challenges and enabling efficient classification.

Experimental setup and results

This section delves into the data used, the processing steps involved, the achieved results, and insights gained from *MemoCMT*. To assess the model's performance, we employed three commonly used datasets for multimodal emotion recognition research: IEMOCAP[41], ESD[42,43], and MELD[44]. All these datasets consist of text and audio recordings simulating realistic conversations between actors and sharing a consistent labeling structure for emotions. We conducted a comprehensive evaluation, analyzing the performance of each model component. This analysis revealed the most effective configurations for emotion recognition. The following paragraph briefly describes the IEMOCAP, ESD, and MELD datasets.

Datasets and training strategy

IEMOCAP Interactive Emotional Dyadic Motion Capture (IEMOCAP)[41] dataset is used for the proposed system evaluation (training and testing) that contains voice utterances and corresponding text transcriptions. To compare the system with existing techniques that used the IEMOCAP corpus with four main emotions: anger, sadness, happiness, and neutral. It should be noted that both improvised and scripted utterances were utilized in this study, including 1103 angry, 1708 neutral, 1084 sad, and 1363 happy (happy and excited) utterances. Because these four emotions are vastly used in literature for systems evaluations, the excited and happy emotions have been combined due to the same feeling based on Plutchik's wheel of emotions.

ESD Emotional Speech Database (ESD)[42,43] comprises over 29 hours of speech data from 10 native English speakers and 10 native Mandarin speakers. It consists of 350 parallel utterances that cover five emotion categories: neutral, happiness, anger, sadness, and surprise. The data was meticulously recorded in a controlled acoustic environment, ensuring accuracy and reliability. An important characteristic of the ESD dataset is its balanced distribution of samples across each emotion class and language. This balanced distribution helps reduce the impact of an imbalanced dataset during training, providing valuable insights into the *MemoCMT*.

MELD Multimodal EmotionLines Dataset (MELD)[44] is an enriched dataset that builds upon the EmotionLines[45] dataset by incorporating audio and visual modalities in addition to text. While preserving the dialogue instances from EmotionLines, MELD introduces enhanced features. With over 1400 dialogues and 13,000 utterances sourced from the Friends TV series, each utterance within a dialogue is categorized with one of seven emotions: anger, disgust, sadness, joy, neutral, surprise, and fear. It is important to note that this dataset originally included three subsets: train, dev, and test. As a result, we used the train set for model development, while the dev and test sets were used to evaluate validation and testing performance.

Model Configuration The model is trained using a sophisticated optimization algorithm called Adam. This optimizer adjusts the model's internal parameters to minimize errors during training. The learning rate, set at 0.0001, controls how drastically these adjustments are made, ensuring precise updates. Additionally, betas (β_1 of 0.9, β_2 of 0.999) help Adam control the decay rates of the moving averages for the gradient and its square, respectively. To enhance the model's convergence, we implement a step learning rate reduction strategy by decreasing the current learning rate by a factor of 0.1 after every 30 epochs, and the model is stopped after 100 cycles (epochs) when it performs best on a separate validation dataset. Additionally, our experiments utilized k=5 for k-fold cross-validation across all datasets. This choice represents a balance between computational efficiency and robust estimation of model performance. Each dataset was divided into five equal subsets, where

four folds were used for training and one for testing. This process was repeated five times, with each fold serving the test set once, ensuring that the model was evaluated on all data points This approach ensures the model captures the most significant relationships within the data while avoiding memorization of the training data itself. *MemoCMT* is trained with a batch size of 1. While training with a batch size of 1 can potentially lead the model to get stuck in local minima, it offers two significant advantages. Firstly, it eliminates the need for padding or trimming audio or text lengths for batching, thereby reducing redundant data. Secondly, it enhances the model's robustness when exposed to varying length inputs during training.

Performance evaluation metrics

W-Acc and *UA-Acc* are two evaluation metrics employed to assess the results. *W-Acc* accuracy is calculated by dividing the number of correct predictions by the total number of samples, which treats all classes equally. However, in the case of IEMOCAP, where there is an imbalance across the emotion classes, *UA-Acc* is also computed. *UA-Acc* assign weights to each class based on the number of samples present in that specific class, providing a more comprehensive evaluation of accuracy that accounts for the varying distribution of samples across different emotions. The formula of *W-Acc* and *UA-Acc* can be defined as follow:

$$W - Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$UA - Acc = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{5}$$

The number of instances correctly predicted as positive by the model is *TP*, while the number of instances incorrectly predicted as positive is *FP*. As a result of the model correctly predicting negative instances as negative, *TN* is the number of negative instances the model correctly predicted as negative. As a result of the model incorrectly predicting positive instances as negative, *FN* is the number of actual positive instances.

Impact of multimodality

Due to the lack of context in the transcripts of the ESD dataset, it cannot benefit from the pre-trained *BERT* model. As a result, we only use four categorical emotions (anger, happiness, neutral, and sadness) from the IEMOCAP dataset for this analysis. Table 1 shows the performances of the unimodal branches of our proposed *MER* system (*SER* for audio and *TER* for text) as well as the multimodal model. We find that using feature fusion of multimodal data significantly improves performance. We achieve exceptional accuracy across all emotion representations using model-level fusion with the primary task (*SER*). Notably, the multimodal model achieves the highest performance with *W-Acc* and *UA-Acc* of 79.25% and 78.92%, respectively, with 8.25-12.69% and 7.80%-13.58% notable improvement compared to the others. The *MER* system benefits from the abstract and general characteristics of the speaker by using pre-trained models for *SER* and *TER*. These models utilize knowledge gained from relevant datasets to generate useful features that help the *MER* accurately predict emotions. The *SER* model captures details about the phonetic and linguistic characteristics of the speech, while the *TER* model focuses on the context and meaning conveyed by the speaker. By learning emotion and speaker characteristics at the same time, we can achieve better results in identifying high-level feature demonstrations, enhancing the *SER* model's efficiency, as mentioned in Table 1.

Impact of feature fusion

To validate the effectiveness and robustness of *MemoCMT*, we assessed the performance of *CMT* with different fusion mechanisms. Table 2 shows the results of *MemoCMT* using *CMT* with different fusion mechanisms on both the IEMOCAP and ESD datasets. For IEMOCAP, the proposed method using *CMT* with CLS achieves the lowest performance, while that with MIN achieves the highest. Specifically, *MemoCMT* using *CMT* with MIN obtains a *W-Acc* of 81.85% and a *UA-Acc* of 81.33%. This trend is observed similarly in the ESD dataset. Notably, *MemoCMT* using *CMT* with MIN achieves the highest performance with a *W-Acc* and *UA-Acc* of 91.84% and 91.93%, respectively. Regarding the IEMOCAP dataset, the model using *CMT* with MIN gains notable improvements of 1.62–2.60% and 1.21–2.41% in *W-Acc* and *UA-Acc*, respectively, compared to the others. In terms of the ESD dataset, it gains improvements of 7.72% in *W-Acc* and 7.72% in *UA-Acc* compared to using *CMT* with CLS. Moreover, these results demonstrate that using the MIN aggregation approach achieves the best performance, indicating that it can be considered to replace the common approaches like CLS and MEAN. These findings suggest that using *CMT* with MIN might reduce the gap between features and modalities, leading to improved performance.

Modality	*W-Acc* (%)	*UA-Acc* (%)
Text	71.10	71.12
Audio	66.56	65.34
Text + Audio[a]	79.25	78.92

Table 1. Impact of modality in our designed *MER* for the IEMOCAP corpus. [a] This experiment employed *CMT* with a class token to ensure comparability with the other unimodal models

Fusion mechanism	IEMOCAP		ESD	
	W-Acc (%)	*UA-Acc* (%)	*W-Acc* (%)	*UA-Acc* (%)
CMT + CLS	79.25	78.92	84.12	84.21
CMT + MEAN	79.47	79.12	90.66	90.71
CMT + MIN	81.85	81.33	91.84	91.93
CMT + MAX	80.23	80.12	91.78	91.86

Table 2. Impact of fusion mechanisms in *MemoCMT* on the IEMOCAP and ESD corpora, respectively. CLS, Class token; MEAN, Mean aggregation; MIN, Min aggregation; MAX, Max aggregation.

Moreover, to delve deeper into the insights of the *CMT*, we employ t-SNE[29] to visualize the learned feature representations of different fusion mechanisms based on the ESD dataset (Fig. 2). We observe that the learned feature representations using *CMT* with CLS (Fig. 2a) and MEAN (Fig. 2b) show significant overlap between the five emotions, while those using MIN (Fig. 2c) and MAX (Fig. 2d) show clearer separation. Interestingly, the learned feature representations in MIN (Fig. 2c) demonstrate significant discriminative ability. These findings are consistent with those presented in Table 2, indicating that the trained model can effectively and robustly classify the five emotional states.

Results and evaluation

The IEMOCAP and ESD datasets were analyzed, and models were trained and tested, with results compared. The confusion matrices depicted in Fig. 3 illustrate the confusion between actual and predicted labels. Figure3a illustrate the confusion matrix of *MemoCMT* using *CMT* with MIN on the IEMOCAP dataset. In this case, *MemoCMT* performs the highest in recognizing the angry emotion and the lowest in recognizing the neutral emotion. On the ESD dataset (Fig. 3b), *MemoCMT* shows the lowest performance in recognizing happiness, while it exceeds 90% accuracy in recognizing the other four emotional states. Interestingly, *MemoCMT* secured better results in recognizing the neutral emotion on the ESD data, with an accuracy of 97.24%.

Moreover, we have plotted the corresponding receiver operating characteristic (*ROC*) curves and calculated the area under the curve (*AUC*) of *MemoCMT* on both the IEMOCAP and ESD datasets (Fig. 4). Figure 4a illustrates the curves for recognizing four emotional states on the IEMOCAP dataset. Notably, it achieves *AUC* values of 0.9738, 0.9426, 0.9611, and 0.9043 for anger, happiness, sadness, and neutral, respectively. Interestingly, *MemoCMT* achieves higher performance in terms of *AUC* on the ESD dataset (Fig. 4b), even though the number of emotions is greater than on the IEMOCAP dataset. It exceeds an *AUC* of approximately 0.9900 for all five emotions.

Comparison and discussion

The performance of our designed *MER* system is evaluated against existing methods, as presented in Table 3. We established a multimodal baseline using an end-to-end pipeline incorporating *CMT* for the *SER* channel and *TER* channel using audio-textual cues. Results in Table 3 demonstrate that our proposed method achieves the highest performance in terms of *UW-Acc* and *W-Acc* on the IEMOCAP dataset. Notably, *MemoCMT* obtains the *UW-Acc* and *W-Acc* values of 81.33% and 81.85%, respectively, which are 5.70–17.83% and 6.55–23.05% higher than those of the other studies. Furthermore, comprehensive details and comparison of the designed model and the ablation study, which utilizes an AI-based architecture for emotion identification, are provided in Tables 1, 2, 3, 4. We have compared several approaches with ours and demonstrated superior accuracy on the IEMOCAP and ESD corpora. Our proposed model exhibits excellent performance, achieving a high recognition rate surpassing existing methods. This notable improvement underscores the robustness and significance of our approach in the field of *MER* that attributed to several factors: (a) The effective integration of speech and text modalities allows for more comprehensive emotion analysis, and (b) The use of advanced architectures like aggregated module and *BERT* capture complex audio and textual data patterns.

A cross-validation technique has been employed extensively to achieve a robust speaker-independent assessment. Using this method, we ensure that the speakers in each fold are different, thereby mitigating potential biases and enhancing the generalization of our results. Specifically, our speech evaluation procedure was structured as follows: (a) The dataset was divided into multiple folds, each containing unique speakers not present in the other folds; (b) During this stage, the decoder underwent training to recognize emotions from input samples and repeated for each fold, with the model learning to recognize emotional patterns across various speakers. (c) The trained classifier must be used to extract feature vectors from the test utterances. By using unseen data in this evaluation phase, we gain a realistic assessment of how well the model can recognize emotions in real-world scenarios. Adopting this rigorous cross-validation methodology ensures that speaker-dependent factors do not artificially inflate our model's performance. This approach provides a more accurate representation of how the system would perform in real-world scenarios with unknown speakers.

Our method improves emotion recognition using multiple data sources, including speech and text. Adding audio data enhances accuracy, and combining speech and text data yields superior results. We also identified the optimal threshold for additional data and highlighted the importance of calibrating the loss function weights. Our approach overcomes data limitations and offers insights into data augmentation strategies and model optimization techniques.

Our model addresses the challenge of partial emotion data accessibility by leveraging multiple sources. This significantly contributes to developing a robust mechanism for emotion recognition using speech and text data.

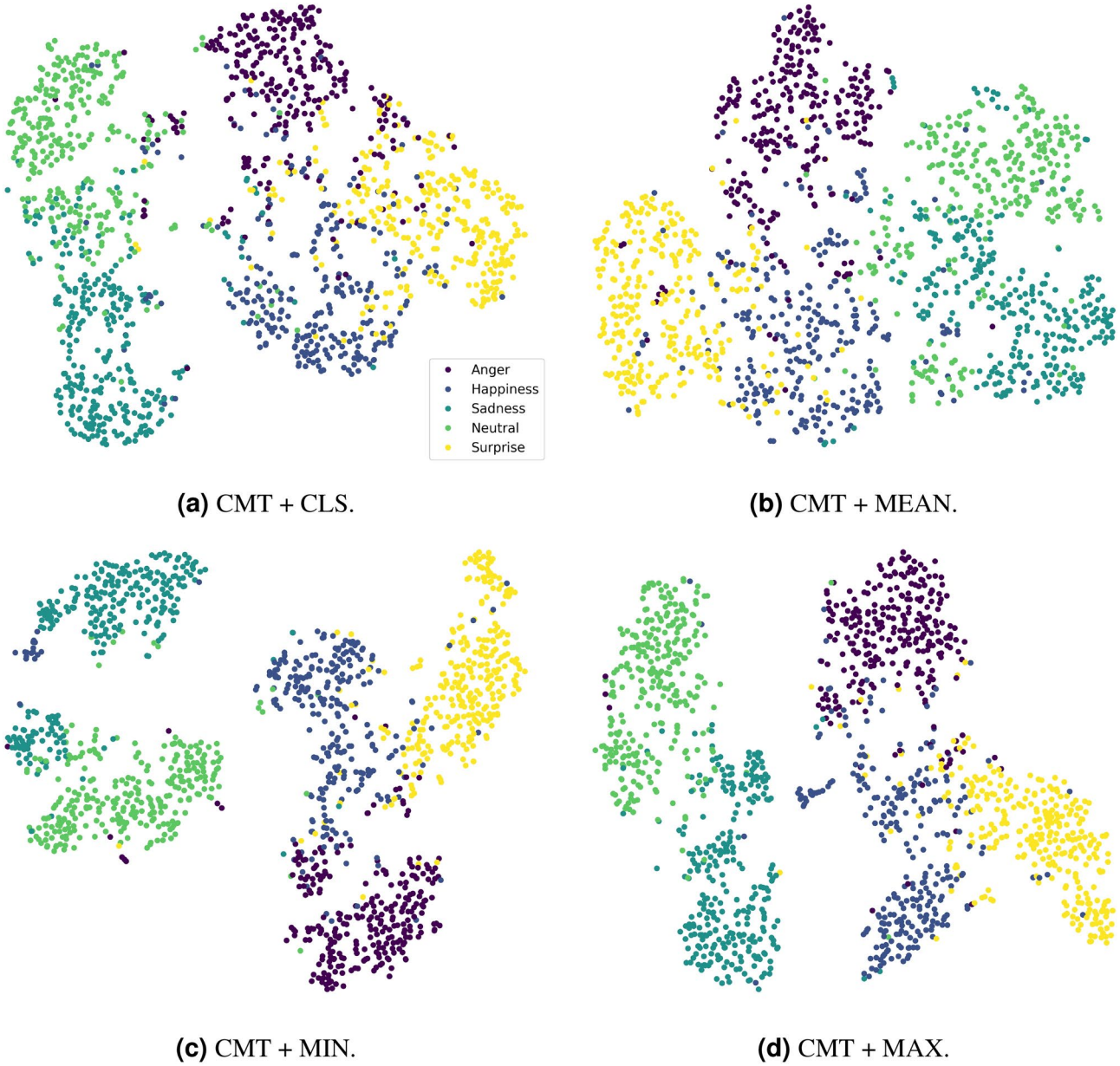
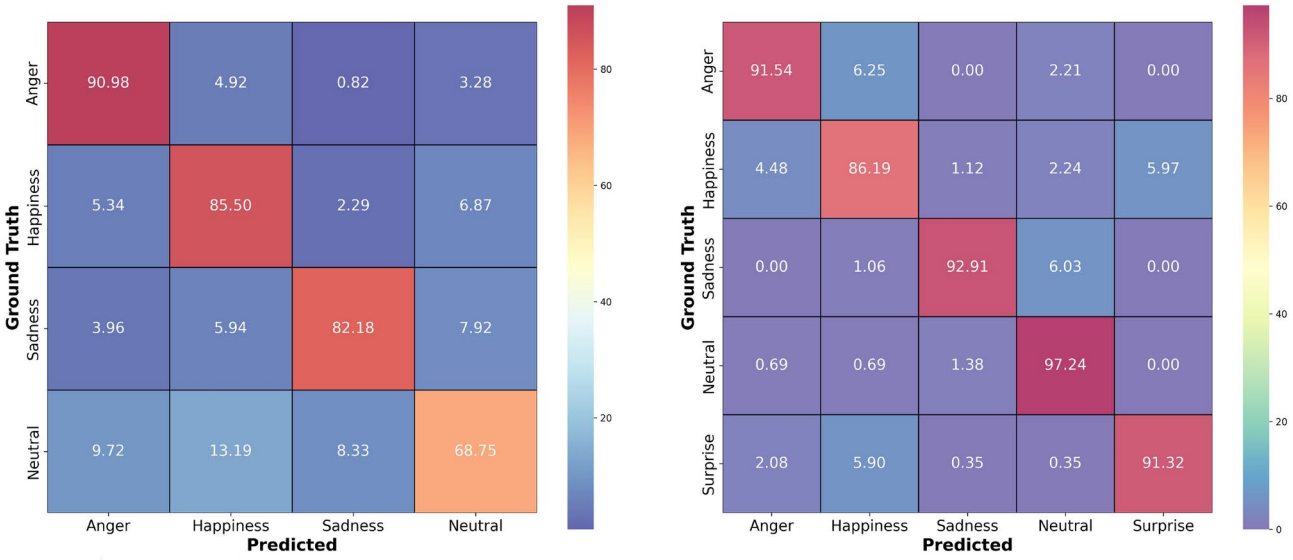


Fig. 2. t-SNE visualizations of *MemoCMT* on the ESD dataset with five emotions (anger, happiness, neutral, sadness, and surprise) using different fusion mechanisms, such as *CMT* with CLS (**a**), *CMT* with MEAN (**b**), *CMT* with MIN (**c**), and *CMT* with MAX (**d**).

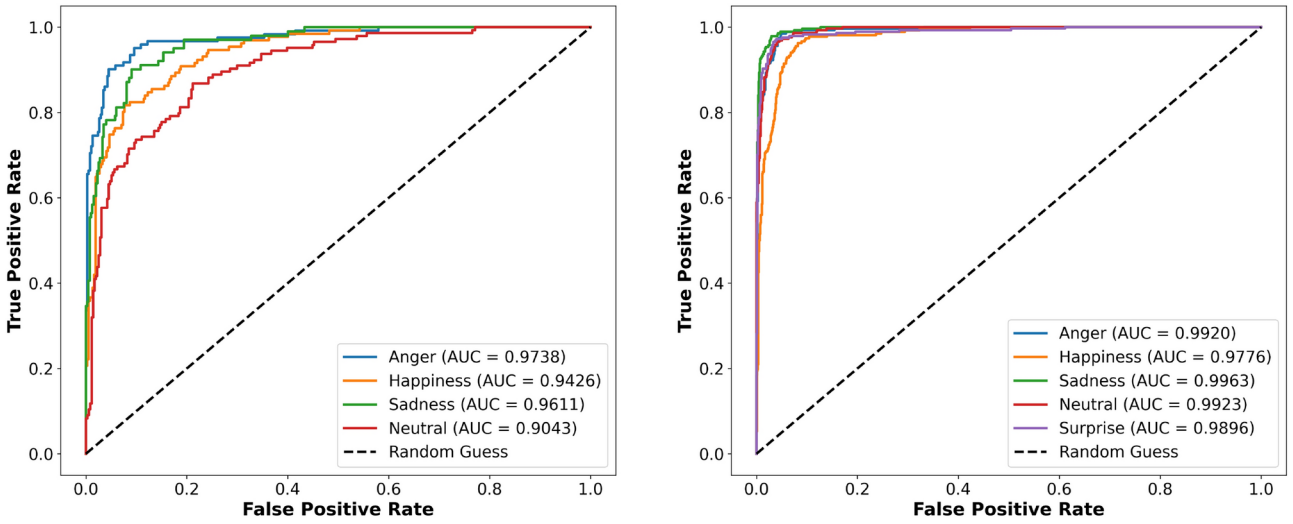
These findings have significant implications for emotion recognition, highlighting the importance of multimodal approaches in capturing the full spectrum of emotional expressions, providing insights into effective data augmentation strategies, and potentially reducing the need for large, manually labeled datasets. They underscore the need to carefully consider model architecture and training procedures, particularly loss function design. Furthermore, our method's ability to overcome partial data accessibility issues suggests its potential applicability in scenarios where complete emotional data may not be available, such as in real-time emotion recognition systems or when dealing with noisy or incomplete datasets. Future work could explore the generalizability of these findings to other datasets and emotion recognition tasks and investigate the potential for incorporating additional modalities (e.g., visual cues) to further enhance the robustness and accuracy of emotion recognition systems.

Table 4 compares the proposed *MER* model's performance on the ESD dataset to baseline methods. The table presents different methods, each associated with its architectural features and the corresponding *UW-Acc* and/or *W-Acc* in percentage. Using a combination of *CMT* with MIN, the proposed *MER* model achieves a *UW-Acc* of 91.93%, outperforming the other methods listed. Our proposed method gains notable improvements of 1.46–3.43% and 1.38–3.34% in *UW-Acc* and *W-Acc*, respectively, compared to the other studies. This table serves as a valuable reference for researchers and practitioners in the *SER* field, demonstrating the proposed model's



(a) IEMOCAP dataset. **(b)** ESD dataset.

Fig. 3. Confusion matrices of *MemoCMT* on both the IEMOCAP dataset (**a**) with four emotions and the ESD dataset (**b**), respectively.



(a) IEMOCAP dataset. **(b)** ESD dataset.

Fig. 4. *AUC-ROC* curves of *MemoCMT* on both the IEMOCAP dataset (**a**) with four emotions and the ESD dataset (**b**), respectively.

effectiveness in recognizing speech emotions in the ESD dataset and showcasing its superiority over existing state-of-the-art approaches.

Case study

Moreover, to validate the effectiveness and robustness of our proposed approach, we assess the performance of *MemoCMT* on the MELD dataset as a case study and compare it with that of the previous study[51] (Table 5 and Fig. 5). It should be noted that we use the same performance metrics, including accuracy (*Acc*), F1-score (*F1*), precision (*Prec*), and recall (*Rec*), reported in the previous study[51] to make a fair comparison.

Table 5 shows the performance of *MemoCMT* on the MELD dataset with both validation and testing phases. It can be seen that the performance varies between different fusion mechanisms. Notably, the proposed method using *CMT* with MEAN achieves the highest performance on the validation phase in terms of *Acc*, with the *Acc*,

Methods	Year	Modality	*UW-Acc%*	*W-Acc%*
Zhang et al.[17]	2021	Audio	–	61.80
Khan et al.[6]	2023	Audio	–	72.75
Mirsamadi et al.[20]	2018	Audio	63.50	58.80
Chen et al [13]	2022	Audio + Text	63.80	–
Tarant et al.[15]	2020	Audio	65.40	–
Zhang et al.[3]	2020	Audio + Text	69.70	68.60
Zhang et al.[3]	2020	Audio + Text	68.70	–
Liu et al.[11]	2023	Audio + Text	69.74	–
Kumar et al.[46]	2021	Audio + Text	75.00	71.70
Chen et al.[47]	2023	Audio + Text	74.30	75.30
Wang et al.[48]	2023	Audio + Text	75.08	72.31
Naderi et al.[19]	2023	Audio	75.63	74.16
Ours	2024	Audio + Text	81.33	81.85

Table 3. Performance comparison of our model with baselines on the IEMOCAP corpus.

Method	Year	Modality	*UW-Acc%*	*W-Acc%*
Zhou et al. [43]	2022	Audio	89.00	–
Yang et al. [49]	2024	Audio	88.50	88.50
Pham et al. [50]	2023	Audio + Text	90.47[a]	90.46[a]
Ours	2024	Audio + Text	91.93	91.84

Table 4. Comparitive analysis of our model with a baseline on ESD corpus. [a]This study employed only four emotions: anger, happiness, neutral, and sadness.

	Validation				Testing			
Fusion mechanism	*Acc*	*F1*	*Prec*	*Rec*	*Acc*	*F1*	*Prec*	*Rec*
CMT + CLS	59.66	55.81	58.41	59.66	61.11	58.40	60.32	61.11
CMT + MAX	60.74	57.05	58.58	60.74	62.95	59.82	60.75	62.95
CMT + MEAN	60.83	57.71	58.18	60.83	62.61	60.14	61.01	62.61
CMT + MIN	60.38	57.99	58.29	60.38	64.18	62.52	63.82	64.18

Table 5. Performance of the proposed method on the additional MELD dataset.

F1, *Prec*, and *Rec* values of 60.83%, 57.71%, 58.18%, and 60.83%, respectively. However, the proposed method using *CMT* with MIN obtains the highest performance in terms of *F1*, with the *Acc*, *F1*, *Prec*, and *Rec* values of 60.38%, 57.99%, 58.29%, and 60.38%, respectively, demonstrating the best model among four different fusion mechanisms. To validate the transferability of our proposed method, we also evaluate *MemoCMT* on the test dataset. Interestingly, the proposed method achieves a similar or even better performance than that on the validation dataset. Importantly, the model using *CMT* with MIN again obtains the highest performance, with the *Acc*, *F1*, *Prec*, and *Rec* values of 64.18%, 62.52%, 63.82%, and 64.18%, respectively.

Furthermore, Fig. 5 shows the comparison between *MemoCMT* and the previous study[51] on the MELD testing dataset. It can be easily observed that our proposed method with different variants achieves better performance in all metrics compared to that of the previous study. Notably, the model with *CMT* and MIN achieves notable improvements of 6.18-33.18% in *Acc*, 7.82–36.22% in *F1*, 5.32-39.52% in *Prec*, and 6.18–33.18% in *Rec*, respectively, compared to all models in the previous study. These results demonstrate that our proposed method is more effective and robust when applied to the MELD dataset, as evidenced by the better performance on four distinct fusion mechanisms.

Conclusion

In this research, we developed *MemoCMT*, an innovative *MER* system to identify emotions from speech signals and text transcripts. *MemoCMT* notably leverages the feature representations extracted by *HuBERT* and *BERT* for speech and text inputs via a novel *CMT* strategy. *MemoCMT* outperformed most current *SER* techniques on the IEMOCAP and ESD datasets. Through ablation analysis, it was proven that using multimodality fusion outperformed unimodality. Additionally, using a novel feature fusion strategy resulted in different performance levels. Importantly, *MemoCMT* using *CMT* with MIN achieved the highest performance on both the IEMOCAP and ESD datasets, with *UW-Acc* and *W-Acc* of 81.33% and 91.93%, and 81.85% and 91.84%, respectively. This

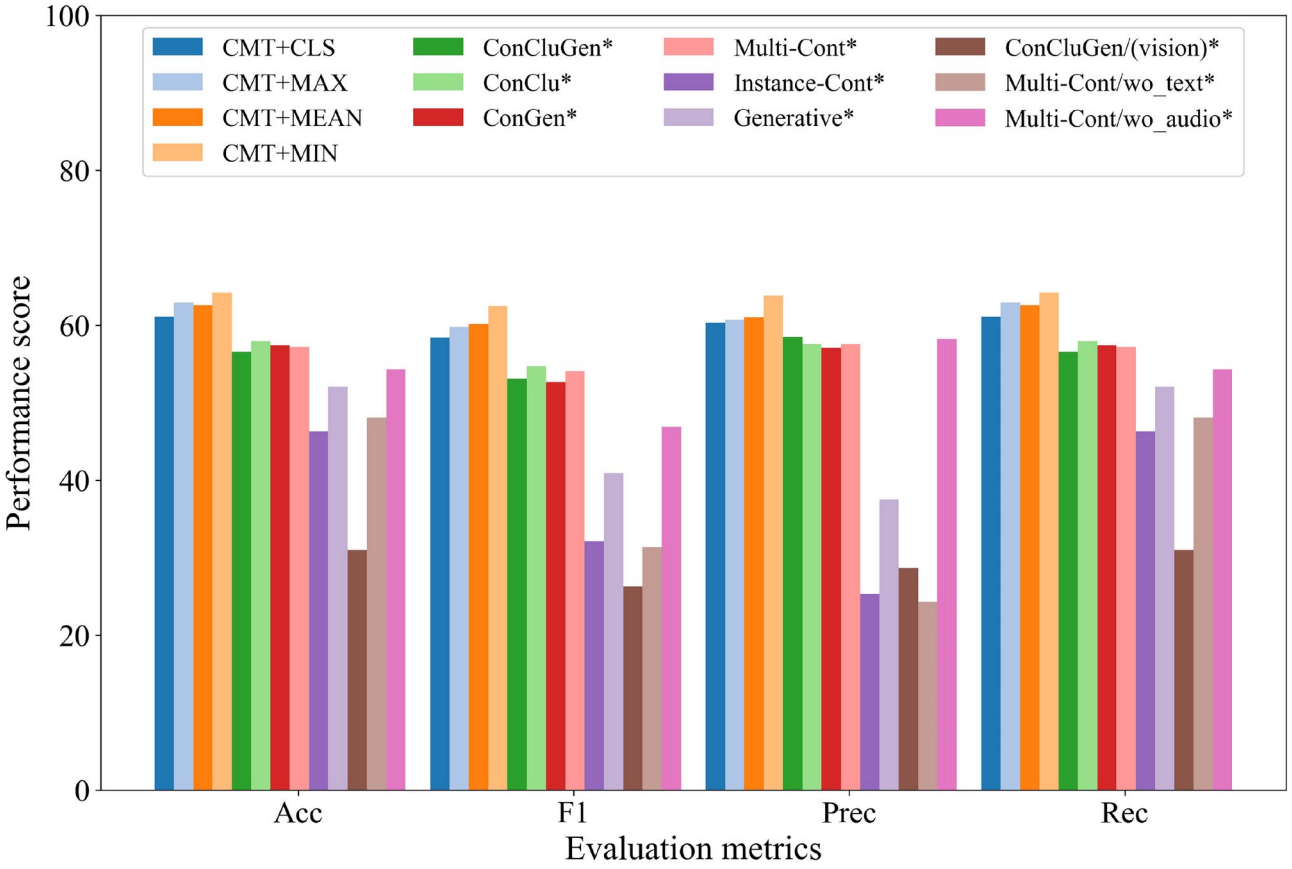


Fig. 5. Performance comparison between *MemoCMT* and the previous study[51] on the same MELD testing dataset. (* denotes the results collected from the previous study[51]).

approach achieved the highest performance on the additional MELD dataset, indicating a need for future studies to explore various aggregation techniques. Such techniques can enhance performance assessment in multimodal analyses using pre-trained models like *BERT* and *HuBERT*. Moreover, we conducted t-SNE visualizations to interpret the effectiveness and robustness of the proposed method.

Although *MemoCMT* is proficient at capturing information from multiple modalities, it has some limitations that must be addressed. Firstly, *MemoCMT* relies on the transformer architecture, which requires high computational resources and is challenging to implement on constrained mobile or IoT devices. Secondly, a drawback of *MemoCMT* is that it requires full modality inputs to be present for predicting emotions, whereas, in real-world scenarios, real-time emotion prediction may necessitate the ability to predict emotions without knowing the maximum length of the input modality or only one of the modality inputs available. In future research, we aim to enhance *MemoCMT* by reducing its complexity by integrating advanced algorithms and tackling the real-time application limitation by exploring techniques like window slicing and length limitation. Moreover, we intend to add new modalities, such as videos and images, and strive to make the system more intelligent and valuable for real-time industrial applications.

Data availibility

The IEMOCAP dataset used in this study is publicly available for research purposes upon request at https://sail.usc.edu/iemocap/. The ESD dataset used in this study is publicly available for research purposes upon request at https://hltsingapore.github.io/ESD/. Please note that this study employed only the English subset of the ESD dataset to develop and evaluate *MemoCMT*. The MELD dataset used in this study is publicly available for research purposes upon request at https://affective-meld.github.io/.

Code availibility

The implementation details of *MemoCMT* are publicly available at https://github.com/tpnam0901/MemoCMT/ for reproducibility purposes.

Received: 26 June 2024; Accepted: 4 February 2025
Published online: 14 February 2025

References

1. El Ayadi, M., Kamel, M. S. & Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recogn.* **44**, 572–587 (2011).
2. Singh, P., Srivastava, R., Rana, K. & Kumar, V. A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowl.-Based Syst.* **229**, 107316 (2021).
3. Zhang, S., Tao, X., Chuang, Y. & Zhao, X. Learning deep multimodal affective features for spontaneous speech emotion recognition. *Speech Commun.* **127**, 73–81 (2021).
4. Tzirakis, P., Zhang, J. & Schuller, B. W. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5089–5093 (IEEE, 2018).
5. Ma, H. et al. A multi-view network for real-time emotion recognition in conversations. *Knowl.-Based Syst.* **236**, 107751 (2022).
6. Mustaqeem, K., El Saddik, A., Alotaibi, F. S. & Pham, N. T. AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network. *Knowl.-Based Syst.* **270**, 110525 (2023).
7. Khan, M., Gueaieb, W., El Saddik, A. & Kwon, S. MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Syst. Appl.* **245**, 122946 (2024).
8. Feng, H., Ueno, S. & Kawahara, T. End-to-end speech emotion recognition combined with acoustic-to-word ASR model. In *INTERSPEECH*, 501–505 (2020).
9. Singh, P., Sahidullah, M. & Saha, G. Modulation spectral features for speech emotion recognition using deep neural networks. *Speech Commun.* **146**, 53–69 (2023).
10. Dai, D. *et al.* Learning discriminative features from spectrograms using center loss for speech emotion recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7405–7409 (IEEE, 2019).
11. Liu, Y., Sun, H., Guan, W., Xia, Y. & Zhao, Z. Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework. *Speech Commun.* **139**, 1–9 (2022).
12. Siriwardhana, S., Reis, A., Weerasekera, R. & Nanayakkara, S. Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition. Preprint at arXiv:2008.06682
13. Chen, Z., Lin, M., Wang, Z., Zheng, Q. & Liu, C. Spatio-temporal representation learning enhanced speech emotion recognition with multi-head attention mechanisms. *Knowl.-Based Syst.* **281**, 111077 (2023).
14. Fayek, H. M., Lech, M. & Cavedon, L. Evaluating deep learning architectures for speech emotion recognition. *Neural Netw.* **92**, 60–68 (2017).
15. Tarantino, L., Garner, P. N., Lazaridis, A. *et al.* Self-attention for speech emotion recognition. In *Interspeech*, 2578–2582 (2019).
16. Yoon, S., Byun, S., Dey, S. & Jung, K. Speech emotion recognition using multi-hop attention mechanism. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2822–2826 (IEEE, 2019).
17. Zhang, R. *et al.* Transformer-based unsupervised pre-training for acoustic representation learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6933–6937 (IEEE, 2021).
18. Xu, M., Zhang, F. & Khan, S. U. Improve accuracy of speech emotion recognition with attention head fusion. In *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 1058–1064 (IEEE, 2020).
19. Naderi, N. & Nasersharif, B. Cross corpus speech emotion recognition using transfer learning and attention-based fusion of wav2vec2 and prosody features. *Knowl.-Based Syst.* **277**, 110814 (2023).
20. Mirsamadi, S., Barsoum, E. & Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2227–2231 (IEEE, 2017).
21. Han, K., Yu, D. & Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. *Interspeech* **2014**, 1–8 (2014).
22. Tang, Y., Hu, Y., He, L. & Huang, H. A bimodal network based on audio-text-interactional-attention with ArcFace loss for speech emotion recognition. *Speech Commun.* **143**, 21–32 (2022).
23. Lin, J., Pan, S., Lee, C. S. & Oviatt, S. An explainable deep fusion network for affect recognition using physiological signals. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2069–2072 (2019).
24. Zhang, J., Huang, X., Yang, L. & Nie, L. Bridge the semantic gap between pop music acoustic feature and emotion: Build an interpretable model. *Neurocomputing* **208**, 333–341 (2016).
25. Fazi, M. B. Beyond human: Deep learning, explainability and representation. *Theory Cult. Soc.* **38**, 55–77 (2021).
26. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, 3145–3153 (PMLR, 2017).
27. Hsu, W.-N. et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291 (2021).
28. Devlin, J., Chang, M. *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1*, 4171–4186, https://doi.org/10.18653/V1/N19-1423 (2019).
29. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 11 (2008).
30. Poria, S., Cambria, E., Bajpai, R. & Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **37**, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003 (2017).
31. Poria, S., Cambria, E., Howard, N., Huang, G. & Hussain, A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* **174**, 50–59. https://doi.org/10.1016/j.neucom.2015.01.095 (2016).
32. Wu, W., Zhang, C. & Woodland, P. C. Emotion recognition by fusing time synchronous and time asynchronous representations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6269–6273 (IEEE, Piscataway, NJ, 2021).
33. Wu, J. et al. A novel silent speech recognition approach based on parallel inception convolutional neural network and mel frequency spectral coefficient. *Front. Neurorobot.* **16**, 971446 (2022).
34. Joshi, A., Bhat, A. & Jain, A. Contextualized GNN based multimodal emotion recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4148–4164 (Stroudsburg, PA, 2022).
35. Lian, Z. *et al.* Context-dependent domain adversarial neural network for multimodal emotion recognition. In *Interspeech*, 394–398 (Cary, NC, 2020).
36. Chen, M. & Zhao, X. A multi-scale fusion framework for bimodal speech emotion recognition. In *Interspeech*, 374–378 (Cary, NC, 2020).
37. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A. & Manocha, D. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1359–1367 (AAAI Press, Menlo Park, CA, 2020).
38. Sun, L., Liu, B., Tao, J. & Lian, Z. Multimodal cross-and self-attention network for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4275–4279 (IEEE, 2021).
39. Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210, https://doi.org/10.1109/ICASSP.2015.7178964 (2015).
40. Xu, J., Sun, X., Zhang, Z., Zhao, G. & Lin, J. Understanding and improving layer normalization. *Advances in Neural Information Processing Systems***32** (2019).
41. Busso, C. et al. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **42**, 335–359 (2008).

42. Zhou, K., Sisman, B., Liu, R. & Li, H. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 920–924 (IEEE, 2021).
43. Zhou, K., Sisman, B., Liu, R. & Li, H. Emotional voice conversion: Theory, databases and esd. *Speech Commun.* **137**, 1–18 (2022).
44. Poria, S. *et al.* MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Korhonen, A., Traum, D. & Màrquez, L. (eds.) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536, https://doi.org/10.18653/v1/P19-1050 (Association for Computational Linguistics, Florence, Italy, 2019).
45. Hsu, C.-C., Chen, S.-Y., Kuo, C.-C., Huang, T.-H. & Ku, L.-W. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018).
46. Kumar, P., Kaushik, V. & Raman, B. Towards the explainability of multimodal speech emotion recognition. In *Interspeech*, 1748–1752 (2021).
47. Chen, W., Xing, X., Xu, X., Yang, J. & Pang, J. Key-sparse transformer for multimodal speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6897–6901 (IEEE, 2022).
48. Wang, Y. et al. Multimodal transformer augmented fusion for speech emotion recognition. *Front. Neurorobot.* **17**, 1181598 (2023).
49. Yang, J. et al. Single-and cross-lingual speech emotion recognition based on WavLM domain emotion embedding. *Electronics* **13**, 1380 (2024).
50. Pham, N. T., Phan, L. T., Dang, D. N. M. & Manavalan, B. SER-Fuse: An emotion recognition application utilizing multi-modal, multi-lingual, and multi-feature fusion. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, 870–877 (2023).
51. Halawa, M. *et al.* Multi-task multi-modal self-supervised learning for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4604–4614 (2024).

Acknowledgements

This work was supported by Déclic, a deep tech startup. Déclic is a social network for outings, meetings, and events, fostering genuine connections and shared activities in the real world (https://declic.net/). We also want to express our gratitude for the AI tools that ensure content clarity throughout.

Author contributions

M.K., P.N.T., and N.T.P.: Conceptualization; data curation; methodology; software; validation; visualization; writing-original draft. A.O., and A.E.S: Conceptualization, Writing-original draft, Supervision. All authors wrote and reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.