# Single-Layered Fusion in Uni-Task Multimodal Speech Emotion Recognition Models

Xuefei Bian, Hao-wei Liang, Tiantian Zhang

Term Paper

in Partial Fulfillment of the Requirements for the Course: Speech Recognition II

MSc Voice Technology

Campus Fryslân

University of Groningen

## Abstract

In this paper, we propose a single-layered fusion uni-tasking Speech Emotion Recognition (SER) multimodal architecture, which reduces training and inference time by 43% and 38%, respectively, compared to a similar multi-layer fusion model, while achieving state-of-the-art accuracy (74.9% on IEMOCAP dataset). The fusion mechanism in the proposed model dynamically integrates acoustic and semantic features using a temporal-aware sigmoid gate, allowing efficient weighting of complementary emotional information while preserving temporal dynamics. The experiments in this paper demonstrate that uni-tasking SER models achieve an optimal trade-off between computational efficiency and accuracy with simpler fusion gates, highlighting the suitability of single-layered fusion uni-tasking SER models for real-time applications.

Key words: speech emotion recognition, multimodality

## 1. Introduction

SER development emphasizes deep learning innovations such as CNNs, RNNs/LSTMs, and Transformers with self-supervised learning. Current advancements prioritize multi-task learning [1], multi-scale feature fusion [2], channel attention [3], and robust techniques [4]. Fusion gates play a pivotal role by enabling dynamic integration of these methodologies, driving improved accuracy and adaptability in SER systems.

Fusion mechanisms in SER aim to combine complementary information from different sources, such as acoustic features, text, or visual data, to enhance emotion recognition accuracy. It dynamically controls information flow in SER models by placing heavier weights on the most relevant features among modalities and suppressing irrelevant features such as noise. It also enhances SER robustness by combining noisy and enhanced features in joint training frameworks, allowing the model to learn from both and perform better under adverse conditions [5]. It has been successfully applied in integrating external language models [6], incorporating conversational-context embeddings [7] into end-to-end speech recognition systems, as well as combining spatial and spectral features in multi-channel speech separation tasks [8].

However, the fusion mechanism often introduces additional computational complexity, thus lengthening training and inference time and limiting usability in real-time applications such as virtual assistants, customer service chatbots, and emergency response systems. Strategies such as lightweight architecture [9], efficient feature extraction [10], and optimized fusion modules [11] have been proposed to help balance accuracy and computational efficiency.

In this paper, we propose a single-layered fusion uni-tasking SER multimodal architecture that significantly improves computational efficiency without compromising accuracy. The architecture employs a temporal-aware sigmoid gate to dynamically integrate acoustic and semantic features, efficiently capturing complementary emotional information while maintaining temporal dynamics. The proposed method uses pre-trained HuBERT [12], wav2vec 2.0 [13], and BERT [14] models for feature extraction, followed by a lightweight fusion mechanism that reduces training time by 43% and inference time by 38% compared to a similar multi-layer fusion model. Experimental validation on the IEMOCAP dataset demonstrates that the model achieves state-of-the-art weighted accuracy (74.9%), outperforming other uni-tasking SER architectures. This highlights the effectiveness of the single-layer fusion gate in achieving an optimal balance between accuracy and computational cost, proving its suitability for real-time applications.

## 2. Related Work

Fusion mechanisms are integrated in both uni-tasking and multi-tasking SER models. Multi-tasking models implement auxiliary tasks include naturalness prediction [15], speaker recognition [16], gender classification [17], and emotion intensity recognition. Uni-tasking approaches, on the other hand, focus on optimizing a single task, typically categorical or dimensional emotion recognition, without leveraging auxiliary tasks.

Uni-tasking models benefit from simpler single-layer fusion gates for a better computational cost/accuracy trade-off, while multi-tasking models require complex fusion gates to integrate auxiliary tasks effectively. We benchmarked our proposed architecture against the following state-of-the-art models.

*2.1 Uni-tasking Model with Multi-layer Fusion Gate*

Gao et al. [18]'s model (CmGI) is one of the top-performing uni-tasking multimodal SER models, achieving a 79.5% weighted accuracy on the IEMOCAP dataset. However, this model requires longer training and inference time compared to the single-layer fusion architecture we propose.

The CmGI model employs a cross-modal gated interaction module, which integrates acoustic features extracted by HuBERT and semantic features extracted by BERT.

The fusion gate consists of a temporal-aware gated mechanism, which preserves the temporal dynamics of input features during fusion. By combining outputs from self-attention and cross-attention layers, the gate adaptively weighs the importance of acoustic and semantic features, enabling the model to effectively handle modality incongruities and improve emotion recognition performance.

*2.2 Uni-tasking Model with Single-layer Fusion Gate*

Yu et al. [19]'s model is the closest to our proposed architecture, as it utilizes a single-layer fusion of cross-attention with a softmax function for both acoustic and semantic features. However, it underperforms our architecture with a weighted accuracy of 69.65% on the IEMOCAP dataset, compared to our 74.9%, due to its limited ability to capture temporal dynamics and fully exploit the complementary information between modalities.

*2.3 Mul-tasking Model*

Ghosh et al. [20]'s MMER model is one of the top-performing multi-tasking multimodal SER architectures, achieving an 81.2% weighted accuracy on the IEMOCAP dataset. This model outperforms many existing SER systems by leveraging a multi-task learning framework, though the inclusion of multiple auxiliary tasks increases its computational complexity compared to simpler architectures.

MMER utilizes a Multimodal Dynamic Fusion Network, which integrates acoustic features extracted by wav2vec-2.0 and textual features extracted by RoBERTa [21]. Through a combination of cross-modal attention mechanisms and auxiliary tasks such as ASR with CTC loss, supervised contrastive learning, and augmented contrastive learning, the architecture captures fine-grained inter-modal emotional representations, addressing modality-specific biases while improving robustness to speaker and semantic variations.

The fusion mechanism employs a cross-modal interaction module composed of multiple Cross-Modal Encoder blocks, which dynamically align and integrate speech and text features. Additionally, an acoustic gate is incorporated to filter redundant speech information, ensuring that only the most relevant acoustic features contribute to the final fused representation. This comprehensive fusion strategy enables MMER to achieve state-of-the-art emotion recognition performance, particularly in challenging multimodal scenarios.

# 3. Proposed Methodology

In the proposed architecture, the acoustic and semantic features are extracted with the pre-trained HuBERT and BERT models, respectively. A cross-modal gated interaction module performs bi-directional cross-attention and temporal-aware gated fusion. The output is fed into a fully connected layer which performs emotion classification task.

*3.1 Acoustic and Semantic Feature Extraction*

In the acoustic modality, we used HuBERT Large (facebook/hubert-large-ls960-ft), which output acoustic hidden states $X_a \in R^{t\times 1024}$, encoding input speech into a sequence of contextualized acoustic representations. In the semantic modality, the audio input is first transcribed through pre-trained wav2vec 2.0 into text and then tokenized with the BERT (bert-base-uncased) tokenizer. The tokenized input is passed through BERT to get the last hidden state ($X_l \in R^{t\prime\times 768}$) from the final transformer layer and subsequently projected to 1024 dimensions to match acoustic representations, resulting in $X'_l \in R^{t\prime\times 1024}$ as the final semantic representations.

*3.2 Cross-attention and Temporal-aware Gated Fusion*

The acoustic and semantic features are fused by a cross-attention gate with a temporal-aware mechanism.

*Cross-Attention*

Query-Key-Value Projections:

$X_a$ attends to $X'_l$: $C_a$ = Attention $(q = X_a, k = X'_l, v = X'_l)$

$X'_l$ attends to $X_a$: $C_l$ = Attention $(q = X'_l, k = X_a, v = X_a)$

Each attention uses linear projections for q, k, and v and follows scaled dot-product attention.

*Temporal Gated Fusion*

The gate is computed using a linear layer followed by a sigmoid activation function. The linear layer has trainable weights and biases that are optimized during training.

$$G = \sigma(W \cdot concat([X, C]) + b)$$

where,

$X \in \{X_a , X'_l\}$ as the original features

$C \in \{C_a , C_l\}$ as the cross-attended features

$W$: trainable weight matrices

$b$: trainable bias vector

$\sigma$: sigmoid activation function

The trainable weights allow the model to dynamically determine how much emphasis should be placed on the original features or the cross - attended features.

The gate mechanism fuses both the acoustic and semantic features.

$$fused_a = X_a \cdot G + C_a \cdot (1 - G)$$
$$fused_l = X'_l \cdot G + C_l \cdot (1 - G)$$

*3.3 Emotion Recognition Classification*

Both $fused_a$ and $fused_l$ are average-pooled along the time dimension to produce a global representation of the acoustic or semantic features for each input sample. This ensures the model handles variable-length inputs by collapsing the time dimension into a fixed-sized representation.

The global representations are further concatenated along the feature dimension (the resulting feature tensor has a dimension of 2048, concatenating 1024 as the fused acoustic feature dimension and another 1024 as the fused semantic feature dimension) to combine the complementary information from both modalities into a single vector, which is finally used for emotion classification.

Emotion classification is undertaken with a fully connected linear layer, which maps the 2048-dimensional features vector to the number of emotion classes (four classes for IEMOCAP: happy, sad, angry, neutral). The output represents the predicted logits of each emotion class.

## 4. Experiments and Results

To testify our hypothesis that a uni-tasking linear fusion gate achieves lower latency while maintaining comparable accuracy compared to a uni-tasking multi-layer fusion gate or a multi-tasking fusion gate, we implemented the proposed architecture alongside the CmGI and MMER models.

*4.1 Experiment 1: Uni-tasking Linear Fusion vs. Uni-tasking Multi-layer Fusion*

In our implementation, we utilized the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [22] for evaluating SER performance. The dataset comprises approximately 12 hours of multimodal data including audio, video, and textual transcriptions, collected from ten American English speakers engaged in dyadic interactions. We adopt the widely used annotation scheme whereby utterances labeled as "excited" are merged with "happy" into a single class, resulting in four final emotion categories: happy, sad, angry, and neutral. The data is split into training, validation, and test sets according to speaker session, with Session 2 held out for testing. The number of utterances closely aligns with prior works, totaling 5,531 samples. As with previous studies, we report performance using unweighted accuracy (UA) and weighted accuracy (WA) metrics, though our codebase implements a single held-out test set rather than 5-fold speaker-independent cross-validation.

Unlike Gao et al. [18], who froze CNN layers after pretraining HuBERT on ASR and SER tasks, our implementation does not explicitly separate ASR and SER stages. Instead, HuBERT and BERT are trained end-to-end simultaneously using only SER loss. Furthermore, instead of relying on ground-truth transcriptions during training, ASR predictions from wav2vec2.0 was used dynamically to simulate real-world inference conditions. Training is conducted using a batch size of 2, learning rate of 1e-5, and cross-entropy loss, with checkpoints saved per epoch.

*4.2 Experiment 2: Multi-tasking Fusion*

Our implementation of the MMER model followed a multimodal approach, integrating speech and text features. The model uses a convolutional-based architecture to extract relevant features and employs multi-task learning for better generalization. The results demonstrate strong performance across emotions, with weighted accuracy of 80.8%.

*4.3 Results*

*4.3.1 Latency Comparison*

The proposed model reduces inference time by up to 72% and training time by up to 68% compared to CmGI, while maintaining comparable accuracy performance. These improvements testified to our hypothesis that for uni-tasking SER models, simpler fusion realizes better computational cost/accuracy trade-off.

**Table 1. Latency Comparison of Single / Multi-layer Fusion for Uni-tasking and Multi-tasking SER**

| Fusion Gate Type | Model Version | Inference Time* | Training Time** | Fusion Formula |
|---|---|---|---|---|
| Uni-tasking | Muti-layer CMGI [] | 209.54 seconds | 874.27 ~ 925.40 seconds | G=σ(W·[SA(x),CA(x)])<br>output=SA(x)×G+CA(x)×(1−G)<br>SA: self-attention; CA: cross-attention |
| | Linear-layer (proposed model) | 59.17 seconds | 272.39 ~ 298.57 seconds | G=σ(W·[acoustic,semantic]+b)<br>output=acoustic×G+semantic×(1−G)<br>b: bias |
| Our Improvement*** | | 38%~72% | 43%~68% | |
| | | | | |
| Multi-tasking | MMER | 59.17 seconds | Around 600 seconds | G=σ(Wg⊤[R;Q]+bg) |

*Inference Time : Total time spent inferring the entire test dataset. This includes the total time spent running all batches together
**Training Time : time for one epoch
***Exact improvement depends on computational baseline, dataset and utterance length, model architecture, and training setup.

*4.3.2 Accuracy Comparison*

Our proposed architecture achieved an accuracy comparable to that of other uni-task fusion multimodal SER models, with slightly lower accuracy compared to selective multi-task fusion models.

**Table 2. Accuracy Comparison between Proposed and State-of-the-art Models**

| Fusion Gate Type | Fusion Gate Architecture | Datasets | Accuracy | Source |
|---|---|---|---|---|
| Uni-tasking | Proposed Model | IEMOCAP | 74.9% | This paper |
| | Lightweight Fusion Model with Time-Frequency Features | IEMOCAP, RAVDESS | IEMOCAP: 74.62%, RAVDESS: 86.11% | Zhang et al., 2024 [23] |
| | Dynamic Convolutional Neural Network + Bi-LSTM | CISIA, Emo-DB, IEMOCAP | CISIA: 59.08%, Emo-DB: 89.29%, IEMOCAP: 71.25% | Lin et al., 2023 [24] |
| | Cross-Attention Fusion (CAF) | Emo-DB, IEMOCAP, RAVDESS | Emo-DB: 97.20%, IEMOCAP: 69.65%, RAVDESS: 81.86% | Yu et al., 2024 [19] |
| | CNN-MGU-Attention | CASIA, Emo-DB | CASIA: 88.90%, Emo-DB: 86.21% | Wang 2023 [25] |
| Multi-tasking | Dual-Attention Mechanism with Conv-Caps and Bi-GRU Features | EMO-DB, IEMOCAP, SITB-OSED | EMO-DB: 90.31% (WA), 87.61% (UA); IEMOCAP: 76.84% (WA), 70.34% (UA) | Maji et al. 2022 [26] |
| | Sparse Cross-Modal Encoder (SCME) + Gated Fusion Module (GF) | IEMOCAP, MELD | IEMOCAP: 82.4%, MELD: 65.0% (WA) | Cui et al., 2024 [27] |
| | Fusion of Facial Expressions and Speech Features | FER 2013, CK+, RAVDESS | 69% | Vardhan et al., 2024 [28] |

Specifically, our proposed infrastructure of HuBERT for SER, wav2vec 2.0 for ASR, and BERT for semantic analysis demonstrated comparable accuracy to other state-of-the-art multimodal models.

**Table 3. Accuracy Comparison between SER Models with Different Pre-train Infrastructures**

| Pre-trained Models | | | SER |
|---|---|---|---|
| **SER** | **ASR** | **Semantic** | **Accuracy** |
| HuBERT | | | 70.8% |
| HuBERT | | | 75.1% |
| LSTM-LAS | | | 63.1% [29] |
| LSTM | LSTM-Attention | | 68.6% [30] |
| LSTM-Attention | DNN-HMM | LSTM-Attention | 76.1% [31] |
| Wav2vec 2.0 | | BERT | 74.2% [32] |
| HuBERT | | BERT | 76.9% |
| | | BERT | 66.9% |
| HuBERT | | BERT | 77.4% |
| HuBERT | Wav2vec 2.0 | BERT | 74.9%* |

*proposed model

## 5. Conclusion and Limitations

The proposed architecture shows overfitting. By epoch 6, training loss drops to 0.1420 and accuracy reaches 95.43%, indicating the model is memorizing training data. However, validation loss rises to 1.0911 and accuracy stagnates at 73%, revealing poor generalization. The IEMOCAP dataset, with 6,412 utterances (~12 hours), limits variability. Lack of regularization (dropout, weight decay, early stopping) and absent data augmentation (noise injection, time stretching) further worsen overfitting. Addressing these issues requires stronger regularization, data augmentation, and training strategies to improve robustness.

In conclusion, this paper presents a single-layered fusion uni-tasking architecture for multimodal SER, achieving a notable balance between computational efficiency and accuracy. While the model shows promise for real-time applications, further work is needed to address overfitting and improve generalization through stronger regularization and data augmentation techniques.

# 6. References


[1] Ma, Y., Di, G., & Wang, W. (2022, November). Advances Research in Speech Emotion Recognition Based on Multi-task Learning. In 2022 4th International Workshop on Artificial Intelligence and Education (WAIE) (pp. 27-31). IEEE.

[2] Chen, M., & Zhao, X. (2020, October). A Multi-Scale Fusion Framework for Bimodal Speech Emotion Recognition. In Interspeech (pp. 374-378).

[3] Hao, Y. (2024, May). Channel Attention Scale Feature Fusion with SwiGLU for Speech Emotion Recognition. In 2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI) (pp. 364-368). IEEE.

[4] Thilakarathne, N. N., Galajit, K., Mawalim, C. O., & Yassin, H. (2024, April). Exploring a Cutting-Edge Convolutional Neural Network for Speech Emotion Recognition. In 2024 5th International Conference on Industrial Engineering and Artificial Intelligence (IEAI) (pp. 110-116). IEEE.

[5] Fan, C., Yi, J., Tao, J., Tian, Z., Liu, B., & Wen, Z. (2020). Gated recurrent fusion with joint training framework for robust end-to-end speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 198-209.

[6] Cabrera, R., Liu, X., Ghodsi, M., Matteson, Z., Weinstein, E., & Kannan, A. (2021). Language model fusion for streaming end to end speech recognition. arXiv preprint arXiv:2104.04487.

[7] Kim, S., Dalmia, S., & Metze, F. (2019). Gated embeddings in end-to-end speech recognition for conversational-context fusion. arXiv preprint arXiv:1906.11604.

[8] Fan, C., Tao, J., Liu, B., Yi, J., & Wen, Z. (2020). Gated Recurrent Fusion of Spatial and Spectral Features for Multi-Channel Speech Separation with Deep Embedding Representations. In INTERSPEECH (pp. 3321-3325).

[9] Yu, S., Meng, J., Zhu, B., & Sun, Q. (2024, August). Cross-Attention Dual-Stream Fusion for Speech Emotion Recognition. In 2024 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM) (pp. 1-6). IEEE.

[10] Song, H., Kang, K., Wei, Y., Zhang, H., & Zhang, L. (2024, May). Speech Emotion Recognition based on Multi COATTENTION Acoustic Feature Fusion. In 2024 Second International Conference on Data Science and Information System (ICDSIS) (pp. 1-4). IEEE.

[11] Cui, L., Zhang, Y., Cui, Y., Wang, B., & Sun, X. (2024). A high speed inference architecture for multimodal emotion recognition based on sparse cross modal encoder. Journal of King Saud University-Computer and Information Sciences, 36(5), 102092.

[12] Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM transactions on audio, speech, and language processing, 29, 3451-3460.

[13] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.

[14] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[15] Atmaja, B. T., Sasou, A., & Akagi, M. (2022). Speech emotion and naturalness recognitions with multitask and single-task learnings. IEEE Access, 10, 72381-72387.

[16] Ma, Y., Di, G., & Wang, W. (2022, November). Advances Research in Speech Emotion Recognition Based on Multi-task Learning. In 2022 4th International Workshop on Artificial Intelligence and Education (WAIE) (pp. 27-31). IEEE.

[17] Latif, S., Rana, R., Khalifa, S., Jurdak, R., & Schuller, B. W. (2022). Multitask learning from augmented auxiliary data for improving speech emotion recognition. IEEE Transactions on Affective Computing, 14(4), 3164-3176.

[18] Gao, Y., Shi, H., Chu, C., & Kawahara, T. (2024). Speech Emotion Recognition with Multi-level Acoustic and Semantic Information Extraction and Interaction. In Proc. Interspeech 2024 (pp. 1060-1064).

[19] Yu, S., Meng, J., Zhu, B., & Sun, Q. (2024, August). Cross-Attention Dual-Stream Fusion for Speech Emotion Recognition. In 2024 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM) (pp. 1-6). IEEE.

[20] Ghosh, S., Tyagi, U., Ramaneswaran, S., Srivastava, H., & Manocha, D. (2022). Mmer: Multimodal multi-task learning for speech emotion recognition. arXiv preprint arXiv:2203.16794.

[21] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[22] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42, 335-359.

[23] Zhang, P., Li, M., Zhao, H., Chen, Y., Wang, F., Li, Y., & Zhao, W. (2024, May). Lightweight fusion model with time-frequency features for speech emotion recognition. In 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 3017-3022). IEEE.

[24] Lin, Z., Hu, Z., & Zhu, K. (2023). Speech emotion recognition based on dynamic convolutional neural network. Journal of Computing and Electronic Information Management, 10(1), 72-77.

[25] Wang, Y. (2023, June). Speech emotion recognition based on CNN-MGU-attention. In International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2023) (Vol. 12707, pp. 1019-1025). SPIE.

[26] Maji, B., & Swain, M. (2022). Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with conv-caps and bi-gru features. Electronics, 11(9), 1328.

[27] Cui, L., Zhang, Y., Cui, Y., Wang, B., & Sun, X. (2024). A high speed inference architecture for multimodal emotion recognition based on sparse cross modal encoder. Journal of King Saud University-Computer and Information Sciences, 36(5), 102092.

[28] Vardhan, J. V., Chakravarti, Y. K., & Chand, A. J. (2024, July). Deep Learning-Based Emotion Recognition by Fusion of Facial Expressions and Speech Features. In 2024 2nd World Conference on Communication & Computing (WCONF) (pp. 1-6). IEEE.

[29] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, “Speech representation learn-ing for emotion recognition using end-to-end asr with factorized adaptation.” in Proc. INTERSPEECH, 2020, pp. 536–540.

[30] H. Feng, S. Ueno, and T. Kawahara, “End-to-end speech emotion recognition combined with acoustic-to-word asr model.” in Proc. INTERSPEECH, 2020, pp. 501–505.

[31] J. Santoso, T. Yamada, S. Makino, K. Ishizuka, and T. Hiramura, “Speech emotion recognition based on attention weight correction using word-level confidence measure.” in Proc. INTERSPEECH, 2021, pp. 1947–1951.

[32] L. Bansal, S. P. Dubagunta, M. Chetlur, P. Jagtap, and A. Gana-pathiraju, “On the Efficacy and Noise-Robustness of Jointly Learned Speech Emotion and Automatic Speech Recognition,” in Proc. INTERSPEECH, 2023, pp. 1863–1867.