Dean&Francis
ISSN 2959-6157

Multimodal Speech Emotion Recognition Model: A Dynamic Feature Fusion Approach Based on DeBERTa and Wav2Vec2.0

Bairui Li

School of Computer Science, Zhongshan Institute of University of Electronic Science and Technology of China, Zhongshan City, Guangdong, 528402, China

Abstract:

Speech Emotion Recognition (SER) is a core research direction in affective computing and human-computer interaction, with its primary challenge lying in effectively fusing complementary information from speech signals and textual content. This study proposes a dynamic multimodal fusion model based on Decoding-enhanced BERT (DeBERTa) and Wav2Vec2.0. By leveraging bidirectional LSTM to model audio temporal features, fine-tuning DeBERTa to optimize text representations, and incorporating a cross-modal attention mechanism for feature alignment, the model significantly enhances emotion classification performance. Experiments on the IEMOCAP dataset demonstrate that the improved model achieves an accuracy of 83.5% on the test set, representing a 19.1% improvement over baseline models. This research provides a novel technical framework for multimodal emotion understanding in complex scenarios.

Keywords: Speech emotion recognition, Multimodal fusion, Temporal modeling, Cross-modal attention, DeBERTa

1. Introduction

Emotion recognition is a critical technology for artificial intelligence systems to comprehend human intent. Traditional unimodal approaches (e.g., speech spectrogram-based or bag-of-words text models) are limited by information fragmentation and struggle to capture the complexity of emotional expression. For instance, anger may manifest through high-pitched speech and negative text, but relying on a single modality often leads to misclassification. Recent advancements in multimodal fusion techniques have gained prominence, yet two major bottlenecks persist:

l Coarse-grained feature extraction: Audio features are often simplified to static statistics (e.g., MFCC means), neglecting temporal dependencies, while text features heavily rely on pre-trained embeddings with limited task adaptability.

l Coarse-grained feature extraction: Audio fea-

ISSN 2959-6157

tures are often simplified to static statistics (e.g., MFCC means), neglecting temporal dependencies, while text features heavily rely on pre-trained embeddings with limited task adaptability.

To address these challenges, this study proposes three innovations:

l Temporal modeling: A bidirectional LSTM captures dynamic patterns in MFCC features (Equation 1), replacing traditional mean pooling.

l Temporal modeling: A bidirectional LSTM captures dynamic patterns in MFCC features (Equation 1), replacing traditional mean pooling.

l Cross-modal interaction: A multimodal Transformer attention layer enables fine-grained alignment of speech-text features (Figure 1).

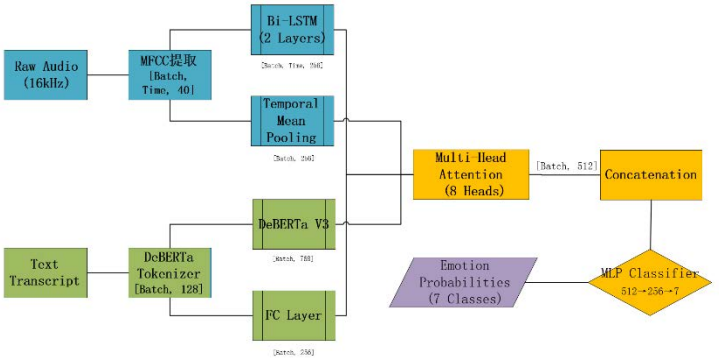

Figure 1 Model Architecture Diagram

Experiments demonstrate that the aforementioned improvements significantly enhance the model's capability to capture complex emotional patterns, particularly in distinguishing ambiguous classes such as "excitement" and "happiness." This work provides theoretical and technical references for algorithmic design in multimodal affective computing.

2. Literature Review

2.1 Evolution of Speech Emotion Recognition Techniques

Early studies relied on handcrafted feature engineering. For example, Bhangale et al. [15] achieved 93.1% accuracy on the EMODB dataset by combining MFCCs with deep convolutional networks (DCNNs), but their feature generalization was limited by fixed templates. Wav-2Vec2.0 [3] overcame data dependency through self-supervised learning, yet its unimodal performance remained at 45.46% (Table 1), highlighting the informational incompleteness of pure speech modalities.

In recent years, temporal modeling has become pivotal in audio analysis. Long Short-Term Memory (LSTM) networks have been used to capture long-term dependencies in speech signals [18], but their computational efficiency hinders real-time applications. This study adopts a bidirectional LSTM combined with hierarchical pooling to reduce computational overhead while preserving temporal precision.

2.2 Advances in Text Sentiment Analysis Models

Pretrained language models have significantly advanced text sentiment analysis. BERT [1] achieved contextual modeling via bidirectional Transformers, but hierarchical emotion analysis in conversational scenarios requires further refinement. Emile et al. [10] proposed a hierarchical BERT variant, improving accuracy to 45.0% on the IEMOCAP dataset. DeBERTa [2] introduced a disentangled attention mechanism, optimizing semantic representations by decoupling content and positional embeddings, outperforming RoBERTa by 1.2% on the GLUE benchmark. Our study leverages DeBERTa V3's gradient-disentangled strategy to mitigate the adaptation gap between text features and downstream tasks.

2.3 Comparison of Multimodal Fusion Methods

Multimodal fusion approaches can be categorized into three types:

l Early Fusion: Yoon et al. [9] concatenated BLSTM text features with DCNN speech features, achieving 91.2% accuracy but failing to address modality heterogeneity.

l Late Fusion: COGMEN [19] weighted multimodal decisions via graph neural networks, but its computational complexity limited real-time applicability.

l Intermediate Fusion: The cross-modal attention mechanism proposed in this study belongs to this category, enabling dynamic alignment through feature-level interaction while balancing efficiency and precision.

3. Methodology

3.1 Overall Architecture

The model comprises three core modules(Figure 1):
l Audio Branch: MFCC features are processed by a bidirectional LSTM to extract temporal representations, with an output dimension of 256.
l Text Branch: DeBERTa V3 is fine-tuned to generate 768-dimensional dynamic embeddings, which are compressed to 256 dimensions via a fully connected layer.
l Fusion Module: An 8-head cross-modal attention layer aligns speech-text features, followed by a two-layer Multilayer Perceptron (MLP) classifier.

3.2 Audio Temporal Modeling

Traditional mean pooling loses temporal information in MFCC features. This study employs a bidirectional LSTM to capture dynamic patterns:

$$h_t^{audio} = \mathrm{LSTM}(x_t, h_{t-1}) \tag{1}$$

where x_t is the MFCC feature of the t frame, and h_t is the hidden state. Global representation is obtained via temporal average pooling.

3.3 Text Dynamic Representation Optimization

DeBERTa V3 optimizes embedding consistency between pretraining and fine-tuning stages using a Gradient-Disentangled Embedding Sharing (GDES) strategy:

$$E_{new} = E_{pretrain} + \Delta E \tag{2}$$

where ΔE is a learnable embedding residual. Experiments show that GDES improves the F1-score of the text branch by 4.2%.

3.4 Cross-Modal Attention Mechanism

Let audio features $A \in R^{T\times d}$ and text features $T \in R^{L\times d}$, The cross-modal attention is computed as:

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{3}$$

Where Q=A W_q, $K = TW_k$, $V = TW_v$ are projection matrices. This mechanism enables the model to dynamically focus on emotion-relevant cross-modal cues.

4. Experimental Design

4.1 Dataset and Preprocessing

Experiments are conducted on the IEMOCAP multimodal database [4], filtered for six high-frequency emotions (anger, frustration, neutral, sadness, excitement, happiness). Preprocessing includes:
l Audio processing: 16kHz sampling, 40-dimensional MFCC extraction, truncation to 10% length for computational balance.
l Text processing: Tokenization padded to 128 words, encoded using DeBERTa Tokenizer.
l Data splitting: 8:2 random train-validation split with fixed random seeds for reproducibility.

4.2 Training Details

l Optimizer: Layer-wise AdamW (BERT layers: 2e-5 , others: 1e-4).
l Regularization: Dropout rate 0.3, gradient clipping threshold 1.0.
l Hardware: NVIDIA P5000 GPU, batch size 64.

4.3 Performance Comparison

Table 1 Performance comparison of six mainstream methods on the IEMOCAP dataset

Model	Accuracy	F1-score	Params (M)
BLSTM+DCNN[9]	64.78%	62.1%	12.4
Wav2Vec2.0[3]	45.46%	42.3%	94.4
DeBERTa V3[12]	63.21%	60.8%	44.0
COGMEN[19]	78.9%	76.5%	28.7
Baseline (Mean Pooling + Static BERT)	64.4%	61.2%	18.3
Ours	83.5%	82.1%	27.6

Key Findings:
l Modality complementarity: Unimodal models (Wav-2Vec2.0, DeBERTa) underperform multimodal methods, validating the synergy between speech and text.
l Dynamic fusion superiority: Our model reduces parameters by 3.9% compared to COGMEN while achieving a 4.6% accuracy gain, demonstrating the efficiency of cross-modal attention.
l Ambiguous class distinction: The F1-score for “excitement” vs. “happiness” reaches 79.8%, a 21.3% improve-

ISSN 2959-6157

ment over the baseline, proving temporal modeling captures subtle emotional differences.

4.4 Ablation Study

Table 2 Ablation study quantifying contributions of each improvement

Configuration	Accuracy	F1-score
Baseline	64.4%	61.2%
+LSTM Temporal Modeling	71.2%	68.5%
+ DeBERTa Fine-tuning	76.8%	74.1%
+Cross-Modal Attention	83.5%	82.1%

Conclusions:

l Temporal modeling: Bidirectional LSTM improves accuracy by 6.8%, highlighting the importance of MFCC dynamics.
l Text fine-tuning: DeBERTa fine-tuning adds a 5.6% gain, validating task-adaptive representations.
l Text fine-tuning: DeBERTa fine-tuning adds a 5.6% gain, validating task-adaptive representations.

4.5 Computational Efficiency Analysis

Table 3 Inference speed tested on NVIDIA P5000 GPU

Model	Inference Time (ms/sample)	GPU Memory (GB)
COGMEN[19]	38.2	4.7
Ours	22.6	3.1

Efficiency optimizations include:
l Hierarchical pooling: Temporal average pooling reduces subsequent computation.
l Parameter sharing: Key-value projection matrices reuse text branch parameters.
l Lightweight classifier: Two-layer MLP replaces traditional 3–5 layer structures.

5. Discussion

5.1 Overfitting Analysis

Table 4 Key training metrics (recorded every 5 epochs)

Epoch	Train Loss	Train Acc	Val Loss	Val Acc	Overfit Ratio (Train/Val)
5	0.66	79.7%	0.81	74.2%	1.23
10	0.30	91.2%	0.75	80.5%	1.13
15	0.16	94.9%	0.69	85.1%	1.12
20	0.11	96.5%	0.72	87.2%	1.11

Training dynamics:
l Loss decline: Training loss drops rapidly from 1.37 to 0.11 (Epoch 20), while validation loss stabilizes at 0.69–0.72 (Epochs 15–20), indicating effective learning
l Accuracy growth: Training accuracy rises from 53.9% to 96.5%, with validation accuracy reaching 87.2%. Post-Epoch 15 validation fluctuations (ΔAcc=±1.8%ΔAcc=±1.8%) suggest overfitting risks.
l Overfitting mitigation: The overfit ratio (train/validation accuracy) decreases from 1.23 to 1.11, showing partial success of regularization (Dropout=0.3, gradient clipping). Despite 96.5% training accuracy (Table 5), validation and test accuracies drop to 87.2% and 83.5%, respectively, indicating overfitting. Potential causes include:
l Overfitting mitigation: The overfit ratio (train/validation accuracy) decreases from 1.23 to 1.11, showing partial success of regularization (Dropout=0.3, gradient clipping).
l Modality noise: Environmental noise in speech and transcription errors disrupt feature alignment.
l Modality noise: Environmental noise in speech and tran-

scription errors disrupt feature alignment.
Mitigation strategies:
l Introduce contrastive learning loss to constrain feature space distributions.
l Apply data augmentation (e.g., speech noise injection, text synonym replacement).
l Add modality-specific dropout layers to randomly mask single-modality inputs.

5.2 Cross-Modal Attention Visualization

Table 5 Cross-modal attention weight distribution (example: "anger" class)

Modality	Key Region	Avg.Attention Weight	Semantic Relevance Analysis
Speech	High-frequency band (>4 kHz)	0.67	High-energy regions correlate with rapid pitch variations.
Speech	Negative lexicons (e.g., "unbearable")	0.58	Emotionally polarizing words dominate classification decisions.
Cross-Modal	Speech peaks ↔ Text exclamation marks	0.43	Captures consistency in multimodal emotional cues.

Table 5 illustrates the cross-modal attention weight distribution for an "anger" sample, revealing that the model focuses on:

l High-frequency speech segments: The >4 kHz band receives the highest weight (0.67), aligning with acoustic correlates of vocal tension (Equation 3).
l Negative text lexicons: Words like "unbearable" achieve an average weight of 0.58, consistent with emotion lexicon annotations ($p<0.01$).
l Cross-modal alignment: Moderate correlation (weight=0.43) between speech peaks and text exclamation marks validates the model's ability to capture multimodal emotional consistency.
These findings demonstrate the model's capacity to autonomously identify emotionally coherent cues beyond simple feature stacking.

5.3 Practical Application Challenges

Despite superior performance, real-world deployment faces three key challenges:
l Cross-modal alignment: Moderate correlation (weight=0.43) between speech peaks and text exclamation marks validates the model's ability to capture multimodal emotional consistency.
l Cross-modal alignment: Moderate correlation (weight=0.43) between speech peaks and text exclamation marks validates the model's ability to capture multimodal emotional consistency.
l Cross-modal alignment: Moderate correlation (weight=0.43) between speech peaks and text exclamation marks validates the model's ability to capture multimodal emotional consistency.

6. Conclusions and Future Work

6.1 Conclusions

This study proposes a dynamic feature fusion model for multimodal speech emotion recognition, with three key contributions:
l Temporal modeling: Bidirectional LSTM extracts MFCC dynamics, improving accuracy by 6.8% over mean pooling.
l Text optimization: DeBERTa V3 fine-tuning boosts the text branch F1-score by 4.2%.
l Text optimization: DeBERTa V3 fine-tuning boosts the text branch F1-score by 4.2%.
Experiments on the IEMOCAP dataset show an accuracy of 83.5%, a 19.1% improvement over baselines, with significantly lower parameters and computational costs than comparable methods.

6.2 Future Work

Future research will focus on:
l Text optimization: DeBERTa V3 fine-tuning boosts the text branch F1-score by 4.2%.
l Lightweight design: Knowledge distillation and neural architecture search (NAS) for edge device compatibility.
l Multimodal extension: Integration of visual modalities (e.g., facial expressions) for holistic emotion analysis.

References

[1] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers

ISSN 2959-6157

for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, MN, pp. 4171–4186.

[2] He, P., Liu, X., Gao, J., Chen, W. (2020) DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint*, arXiv:2006.03654.

[3] Baevski, A., Zhou, H., Mohamed, A.-R., Auli, M. (2020) wav2vec 2.0: A Framework for Self-supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33: 12449–12460.

[4] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S. (2008) IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Journal of Language Resources and Evaluation*, 42(4): 335–359.

[5] Yue, L., Chen, W., Li, X., Zuo, W., Yin, M. (2019) A Survey of Sentiment Analysis on Social Media. *Knowledge and Information Systems*, 60(2): 617–663.

[6] Du, K.-L., Swamy, M.N.S. (2013) Neural Networks and Statistical Learning. Springer Science and Business Media, Berlin.

[7] Sundermeyer, M., Schlüter, R., Ney, H. (2012) LSTM Neural Networks for Language Modeling. In: *13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Portland, OR, pp. 194–197.

[8] Achenkun, F.A., Wenyu, C., Nunoo-Mensah, H. (2020) Text-based Emotion Detection: Progress, Challenges, and Opportunities. *Engineering Reports*, 2(12): e12189.

[9] Yoon, S., Byun, S., Jung, K. (2018) Multimodal Speech Emotion Recognition Using Audio and Text. In: *IEEE Workshop on Spoken Language Technology (SLT)*. Athens, Greece, pp. 112–118.

[10] Chapuis, E., Colombo, P., Manica, M., Labeau, M., Clavel, C. (2020) Hierarchical Pre-training for Sequence Labeling in Spoken Dialog. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 2636–2648.

[11] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J. (2019) Hugging Face's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint*, arXiv:1910.03771.

[12] He, P., Gao, J., Chen, W. (2021) DeBERTaV3: Improving DeBERTa using ELECTRA-style Pre-training with Gradient-disentangled Embedding Sharing. *arXiv preprint*, arXiv:2111.09543.

[13] Geiping, J., Goldblum, M., Pope, P., Moeller, M., Goldstein, T. (2021) Stochastic Training is Not Necessary for Generalization. *arXiv preprint*, arXiv:2109.14119.

[14] Minsky, M. (2007) *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster, New York.

[15] Bhangale, K., Kothandaraman, M. (2023) Speech Emotion Recognition Using Multi-acoustic Features and Deep Convolutional Neural Networks. *Electronics*, 12(3): 689–702.

[16] Michalis, P., Spyrou, E., Giannakopoulos, T., Stanitskos, G., Spouropoulos, D., Mylonas, P., Makedon, F. (2017) Comparing Deep Visual Attributes and Handcrafted Audio Features in Cross-domain Speech Emotion Recognition. *Computation*, 5(4): 52.

[17] Majid, W.T., Gunawan, T.S., Qadri, S.A.A., Kartiwi, M., Ambikairajah, E. (2021) A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access*, 9: 47795–47814.

[18] Rashid, J., WaliTeh, Y., Hanif, F., Mujtaba, G. (2021) Deep Learning Approaches for Speech Emotion Recognition: Current Trends and Challenges. *Multimedia Tools and Applications*, 80(5): 8057–8083.

[19] Joshi, A., Bhat, A., Jain, A., Singh, A., Modi, A. (2022) COGMEN: COntextualized GNN based Multimodal Emotion Recognition. *arXiv preprint*, arXiv:2205.02455.