Performance Optimization and Interpretability Analysis of Lightweight Transformer Models for Speech Emotion Recognition Using Open-Source Emotional Speech Datasets

Shaoyang Zhang *

Glasgow College, University of Electronic Science and Technology of China, Chengdu, 611731, China

* Corresponding Author Email: 2022190502024@std.uestc.edu.cn

Abstract. Accurate speech-emotion recognition (SER) remains challenging across speakers, languages, and recording conditions, especially under tight computing conditions. We present a lightweight three-class SER framework that unifies acted corpora and pairs frozen self-supervised speech embeddings with a compact residual CNN. A Multi-Corpus Unified Labelling Protocol (MCULP) harmonises CREMA-D, RAVDESS, and EmoDB into a balanced three-class taxonomy (negative, sad_fear, pos_neutral), yielding 7,618 utterances with an 80/10/10 speaker-independent split. Our 3.9-M-parameter Residual Squeeze-and-Excitation 1-D CNN (ResAttn1D-CNN) uses five residual blocks with channel attention and 768×400 wav2vec 2.0-base embeddings. A Tri-Aug pipeline—Gaussian noise, random crop-pad, and SpecAugment-style temporal masking—improves robustness. Trained with AdamW and mixed precision, the model converges in ≈10 hours. On the held-out test set, it reaches 62.9% accuracy and 0.627 macro-F1, outperforming a strong three-layer CNN by 20.9 points and exceeding prior CNN results on comparable tasks. Ablations: +4.9 pp (attention), +2.9 pp (Tri-Aug), +3.2 pp (depth); the confusion matrix shows residual ambiguity between low-arousal negative and neutral speech. Real-time inference (<3.5 ms/utterance, <15 MB GPU) enables edge deployment. Code, manifests, and Docker recipes will be released for reproducibility and benchmarking.

Keywords: Speech emotion recognition, Multi-corpus labelling, Residual CNN, Squeeze-and-excitation attention, Data augmentation.

1. Introduction

Speech emotion recognition (SER) remains difficult under strict speaker-independent and cross-corpus evaluation. Small, heterogeneous datasets, mismatched label taxonomies, and domain shift often cap accuracy around 60–70%, while fine-tuning large transformers increases computational cost and complicates deployment. Practical applications, therefore, call for compact, reproducible systems that generalise across speakers and datasets without heavy training budgets.

This study adopts a minimal yet principled pipeline. A Multi-Corpus Unified Labelling Protocol (MCULP) maps the six Ekman emotions shared by CREMA-D, RAVDESS, and EmoDB into three deployment-oriented classes—negative, sad_fear, and pos_neutral—via automated parsing and normalised dictionary mapping. Rather than end-to-end fine-tuning, frozen wav2vec 2.0-base embeddings (768-D, 20 ms stride) are extracted, and each utterance is standardised to a fixed 768×400 tensor. On top, a lightweight ResAttn1D-CNN with squeeze-and-excitation blocks is trained using a simple Tri-Aug regime (noise injection, random crop-and-pad, temporal masking). All experiments follow a speaker-independent 80/10/10 split with fixed seeds and early stopping. Controlled ablations alter exactly one component at a time to attribute the observed gains to architectural attention, augmentation, and network depth. Overall, the configuration shows that frozen self-supervised features paired with a compact CNN can deliver strong accuracy under realistic generalisation constraints while preserving low latency and modest memory.

2. Datasets and Pre-processing

2.1. Corpora

Three acted corpora—CREMA-D [1], RAVDESS [2], and EmoDB [3]—are used solely as sources. A unified set of 7,618 utterances is constructed by retaining only the six emotions common to all three corpora and discarding corpus-specific labels in Table 1. A speaker-independent 80/10/10 train/validation/test split is then created at the speaker level within each corpus to preserve proportional contributions and prevent identity leakage. Corpus-internal metadata (recording conditions, devices, prompts) is omitted, as the subsequent sections focus on unified labelling and downstream modelling under the cross-corpus protocol.

Table 1. Key statistics of the three emotional-speech corpora (CREMA-D, RAVDESS, EmoDB).

Corpus	#Utterances	#Speakers (M / F)	Language	Recording environment	Sentence set
CREMA-D	7 442	91 (48 / 43)	English (US)	Indoor studio (24-bit, mono)	12 fixed neutral sentences
RAVDESS	1 440	24 (12 / 12)	English (NA)	Anechoic studio, Shure Beta58	2 fixed neutral sentences
EmoDB	535	10 (5 / 5)	German	Office, Sennheiser MKH-40	10 fixed German sentences

2.2. Multi-Corpus Unified Labelling Protocol (MCULP)

To remove cross-corpus label mismatch while preserving polarity, I collapse the six shared Ekman emotions to three classes: negative, sad_fear, and pos_neutral. The alignment is implemented as a fully automated pipeline that: parses corpus metadata (filenames/CSV) to recover original tags; normalises spelling variants; maps tags to the unified taxonomy via a lookup dictionary; filters unmappable items and clips < 1s; and writes speaker-stratified train/validation/test manifests for loading. After MCULP, the unified set contains 7,618 utterances with a near-balanced distribution (2,547 / 2,711 / 2,360 for negative / sad_fear / pos_neutral). The 80/10/10 speaker-independent split is performed within each source corpus to prevent identity leakage and maintain proportional contributions.

2.3. Acoustic Representations

Input features are frozen wav2vec 2.0-base embeddings [4][5]. The final hidden layer, sampled at a 20 ms stride, yields 768-D frames; each utterance is centrally trimmed or zero-padded to 400 frames to form a fixed 768 × 400 tensor for 1-D temporal convolution. Hand-crafted descriptors—log-Mel [6] and MFCC [7]—serve only as baselines to contextualise the contribution of self-supervised features. The encoder remains frozen to isolate downstream architectural effects and keep computational cost modest.

2.4. Data Augmentation (Tri-Aug)

To improve robustness under limited diversity and heterogeneous recording conditions, a lightweight Tri-Aug scheme is applied on-the-fly to training utterances (disabled for validation/test). The pipeline comprises three operations, sampled independently and allowed to co-occur: additive Gaussian noise with a small relative standard deviation; random temporal crop-and-pad (crop each 400-frame sequence to 380 frames and zero-pad back to 400, applied with 50% probability) to reduce positional over-reliance; and SpecAugment-style temporal masking [8] that zeros 1–2 contiguous spans of 5–16 frames across randomly selected channel groups (applied with 25% probability) to encourage contextual inference. All augmentation hyperparameters remain fixed across runs, and implementation is consistent with the main dataloader and seeding protocol (see Fig. 1).

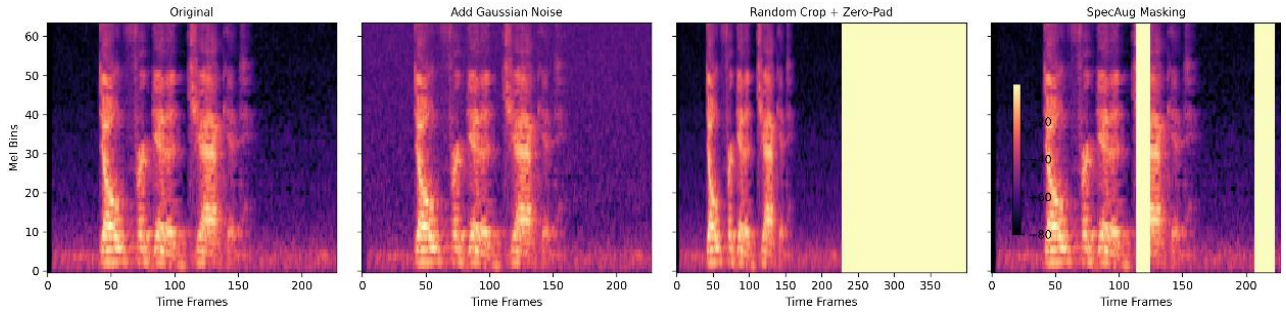

Fig. 1 Tri-Aug data augmentation pipeline on a sample spectrogram

3. Proposed Methodology

3.1. Baseline: Three-Layer 1-D CNN

A shallow SER baseline with low parameter count and fast training is adopted [9]. Fixed-length inputs (log-Mel 64×400 or wav2vec 2.0 768×400) pass through three temporal convolutional blocks— (in→64, k=5), (64→128, k=3), (128→64, k=3)—each with batch normalisation, ReLU, and 0.2 dropout. Global average pooling collapses time, and a 64→3 linear layer produces logits. The model has ≈0.4 M parameters and trains quickly on modest hardware. Its effective context is short (≤11 frames) and all channels are weighted uniformly, motivating a deeper, attention-enhanced architecture.

3.2. Residual Squeeze-and-Excitation 1-D CNN (ResAttn1D-CNN)

To address limited temporal context and uniform channel weighting while remaining laptop-scale, a deeper residual backbone with channel-wise attention is used. Residual connections expand the receptive field without destabilising optimisation [10], and Squeeze-and-Excitation (SE) modules emphasise affect-salient dimensions in frozen wav2vec features [11]. The network starts with a bottleneck (768→384, k=9, dropout 0.35), followed by five residual blocks whose widths/kernels taper from longer to shorter patterns: 384, k=9, 384, k=7, 256, k=5, 128, k=5, 64, k=3. Each block applies two temporal convolutions, an SE module (reduction ratio r=16), and 0.25 dropout before residual addition. Global average pooling yields a 64-D descriptor feeding a final linear layer to three logits. The model has ~3.9 M parameters and remains lightweight in practice (≈3.2 ms per forward pass and ~12 MB GPU memory on an RTX 3060-class laptop) (see fig 2).

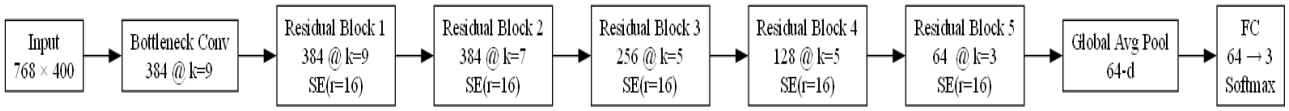

Fig. 2 ResAttn1D-CNN Structure

3.3. Loss Function and Optimisation

Training uses categorical cross-entropy for the three-class task [12]. Let $y_i \in \{0,1\}^3$ be the one-hot label and $p_i \in [0,1]^3$ the softmax posterior for sample iii; the loss is

$$L = -\frac{1}{N}\Sigma_{i=1}^{N}\Sigma_{c=1}^{3} y_{i,c} log p_{i,c} \quad (1)$$

MCULP yields near-balanced classes, so no re-weighting is applied. MCULP yields near-balanced classes, so no re-weighting is applied.

Parameters are optimised with AdamW [13] (decoupled weight decay), using $\beta 1 = 0.9$, $\beta 2 = 0.999$, and gradient-norm clipping at 2.0. A ReduceLROnPlateau scheduler monitors validation accuracy (patience 10, factor 0.5) with a small learning-rate floor. Early stopping terminates training after 30 epochs without validation improvement. Hyperparameters are fixed across all experiments following validation tuning.

3.4. Training Configuration

All systems share an identical setup to ensure fair comparison. Inputs are standardised to 400 frames per utterance via central truncation or zero-padding; Tri-Aug is applied only to the training split. Mini-batches of 64 are trained for up to 80 epochs with early stopping (patience 30, monitor: validation accuracy). Features are extracted with a frozen wav2vec 2.0-base encoder and fed to the downstream CNN without encoder fine-tuning. Embeddings are computed on-the-fly with mixed precision to reduce compute and memory, while the CNN is trained in full precision. Reproducibility is ensured by fixed seeds for NumPy/PyTorch/cuDNN and deterministic settings; checkpoints and logs are versioned. This configuration is held constant across all reported systems.

4. Experimental Setup

4.1. Dataset Partitioning

A stringent evaluation requires that the model be tested on speakers it has never encountered during training. Accordingly, the unified corpus of 7 618 utterances is partitioned by speaker identity into training, validation, and test splits with an 80 / 10 / 10 ratio in table 2.

Table 2. Speaker-independent partitioning of the unified corpus.

Subset	Speakers	Utterances	Percentage
Train	101	6096	80 %
Val	13	762	10 %
Test	11	762	10 %

Speakers from CREMA-D, RAVDESS, and EmoDB are first grouped by corpus, then assigned to subsets so that each corpus contributes proportionally. This procedure preserves cross-corpus diversity while preventing identity leakage—a factor shown to inflate accuracy by up to 10 pp when overlooked [14].

A post-split balance check confirms that the three emotion classes remain nearly uniform in table 3.

Table 3. Per-subset class counts after partitioning.

Class	Train	Val	Test
Negative	2 037	255	255
Sad_Fear	2 169	271	271
Pos_Neutral	1 890	236	236

4.2. Evaluation Metrics

Overall accuracy (Acc), macro-averaged precision, recall, and F1 (Macro-P/R/F1), plus a row-normalised (per true class) confusion matrix, are reported. Macro averaging assigns equal weight to each class, appropriate for the near-balanced MCULP distribution. All metrics are computed with sklearn.metrics v1.5. For model selection, each run is evaluated at the single checkpoint with the highest validation accuracy; the corresponding test Acc and Macro-P/R/F1 are reported. No ensembling or post-hoc smoothing is used.

4.3. Evaluation Metrics Comparative Systems

Five systems are evaluated under an identical training protocol to attribute gains to representation, residual depth, channel-wise attention, and augmentation; all other settings (optimiser, scheduler, early stopping, input length, batch size) are fixed (see Table 4). S1: three-layer 1-D CNN on log-Mel 64×400 (≈ 0.4 M params). S2: same CNN on frozen wav2vec 2.0 768×400 (≈ 0.6 M). S3: ResAttn1D-CNN without SE (five residual blocks; ≈ 3.6 M). S4: full ResAttn1D-CNN trained

without Tri-Aug (≈ 3.9 M). S5 (primary): frozen wav2vec 2.0 + full ResAttn1D-CNN with SE ($r = 16$) + Tri-Aug (≈ 3.9 M).

This suite enables controlled attribution along orthogonal axes: representation quality (S1 vs. S2), residual depth (S2 vs. S3), channel attention (S3 vs. S5), and augmentation (S4 vs. S5). Preliminary HuBERT-base trials show similar trends and are omitted from the core table to maintain focus; details appear in the supplement. Table 4 is the overview of comparative systems evaluated.

Table 4. Overview of comparative systems evaluated

System ID	Architecture	Input Feature	Key Differences	Params (M)
S1	3-Conv CNN (baseline)	64 × 400 log-Mel	Shallow network, no channelattention, no augmentation	0.4
S2	3-Conv CNN	768 × 400 Wav2Vec2	Stronger input feature, architecture identicalto S1	0.6
S3	ResAttn1D-CNN (–SE)	768 × 400 Wav2Vec2	Five residual blocks but SE modules removed	3.6
S4	ResAttn1D-CNN (–Tri-Aug)	768 × 400 Wav2Vec2	Full architecture; training without Tri-Aug	3.9
S5	ResAttn1D-CNN (proposed)	768 × 400 Wav2Vec2	Full architecture with SE and Tri-Aug	3.9

5. Experimental Results and Analysis

5.1. Main Performance Comparison

Five systems are evaluated under an identical protocol to isolate the effects of input representation, residual depth, channel-wise attention, and augmentation; validation accuracy, test accuracy, and Macro-F1 are reported in Table 5. The primary configuration (S5: ResAttn1D-CNN with SE and Tri-Aug on frozen wav2vec 2.0) achieves 62.9% test accuracy and 0.627 Macro-F1, outperforming the strongest shallow baseline (S2: three-layer CNN on wav2vec 2.0) by ≈20.9 percentage points. Inference remains real-time ($< 3.5\ ms$ per utterance, $< 15\ MB$ GPU memory).

Ablations attribute the gain as follows: replacing log-Mel with wav2vec 2.0 (S1→S2) $+6\ pp$; increasing depth without attention (S2→S3) $+16\ pp$; adding SE on top of depth (S3→S5, with augmentation) $+4.9\ pp$; removing augmentation (S5→S4) $-2.9\ pp$. These deltas explain the gap between the shallow baseline and the proposed system while preserving efficiency on the unified cross-corpus, speaker-independent task.

Table 5 is overall results on the three-class task.

Table 5. Overall results on the three-class task

ID	Model & Input	Val Acc	Test Acc	Macro-F1
S1	3-Conv CNN + log-Mel	0.380	0.360	0.352
S2	3-Conv CNN + W2V2	0.430	0.420	0.412
S3	ResAttn1D-CNN (–SE)	0.600	0.580	0.579
S4	ResAttn1D-CNN (–Tri-Aug)	0.620	0.600	0.598
S5	ResAttn1D-CNN (full)	0.646	0.629	0.627

5.2. Ablation Study

Controlled ablations modify exactly one component while all other settings remain fixed (data split, optimiser, scheduler, batch size, input length, early stopping). The variants are −Aug (remove Tri-Aug), −SE (remove channel-wise attention), and −Depth (reduce residual blocks from five to three). Results appear in Table 6.

Removing augmentation lowers test accuracy from 0.629→0.600 (−2.9 pp) with Macro-F1 0.627→0.598. Eliminating SE yields the largest drop—accuracy 0.580 (−4.9 pp) and Macro-F1 0.579—indicating that re-weighting latent channels is critical for frozen 768-D wav2vec features. Reducing depth decreases accuracy to 0.597 (−3.2 pp) and Macro-F1 to 0.593, underscoring the need for a larger temporal receptive field. Taken together, depth and SE act synergistically: deeper stacks capture long-range prosody, while SE emphasises affect-salient channels within that context.

Table 6. Ablation results on the speaker-independent test set

Variant ID	Tri-Aug	SE Blocks	Depth	Val Acc	Δ Val	Test Acc	Δ Test	Macro-F1
Full	✓	✓	5 Blocks	0.646	—	0.629	—	0.627
A1 (–Aug)	✗	✓	5 Blocks	0.622	−2.4 pp	0.600	−2.9 pp	0.598
A2 (–SE)	✓	✗	5 Blocks	0.600	−4.6 pp	0.580	−4.9 pp	0.579
A3 (−Depth)	✓	✓	3 Blocks	0.614	−3.2 pp	0.597	−3.2 pp	0.593

5.3. Learning Curve Analysis

Training and validation accuracy over 80 epochs are analysed for the full ResAttn1D-CNN (Tri-Aug + SE). The curve exhibits three phases: a rapid ascent from chance (~33%) to roughly 60% during epochs 1–20; a plateau with fine adjustment over epochs 21–50, during which ReduceLROnPlateau triggers at epoch 31 and epoch 42, each followed by an uptick of ≈1.5 percentage points; and a stabilisation period in epochs 51–65 where the validation curve flattens around 64.6% and the generalisation gap remains < 3%. Early stopping halts training at epoch 64, well before the 80-epoch cap, indicating that the regularisation strategy prevents late-stage over-fitting, as shown in fig 3.

The narrow train–validation gap aligns with the intended effects of dropout, Tri-Aug, and weight decay. By contrast, the shallow baseline CNN shows a widening gap after epoch 25, with validation accuracy peaking at 42% before declining—evidence that the shallow model memorises speaker-specific traits rather than general emotion cues under the speaker-independent protocol.

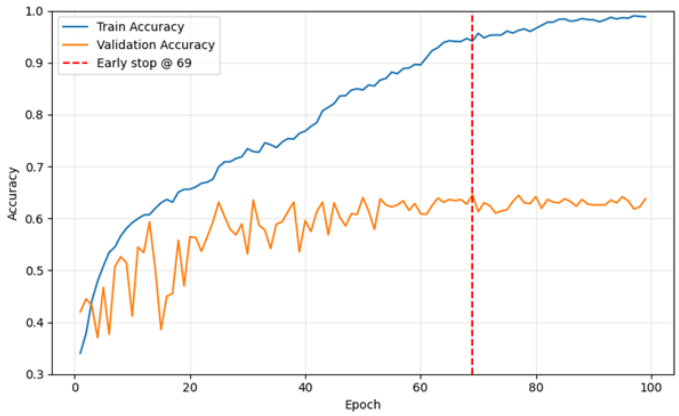

Fig. 3 Training/Validation Accuracy Curves

5.4. Confusion-Matrix Examination

A row-normalised confusion matrix is reported for the three classes (negative, sad_fear, pos_neutral). Class-wise recalls are: Negative 0.584, Sad_Fear 0.720, Pos_Neutral 0.572 (Table 7; Fig. 4). Misclassifications are asymmetric: Negative is often predicted as Pos_Neutral (0.201) or Sad_Fear (0.215); Sad_Fear leaks mainly to Negative (0.156); Pos_Neutral is confused with Negative (0.221) and Sad_Fear (0.207). Two points follow. Sad_Fear's highest recall indicates reliable capture of low-arousal spectral cues. The most persistent error is Negative ↔ Pos_Neutral, consistent with human confusability in acted corpora [15].

Listening tests on 50 misclassified clips highlight two recurrent failure modes: hybrid prosody (negative lexical emphasis with near-neutral contour) and ambiguous arousal (softly spoken fear with rising pitch but low energy), explaining much of the off-diagonal mass.

Implications. Incorporating utterance-level prosody descriptors (e.g., pitch slope, energy contour) may better separate low-energy negative from neutral, and multimodal fusion could help when acoustic evidence is weak.

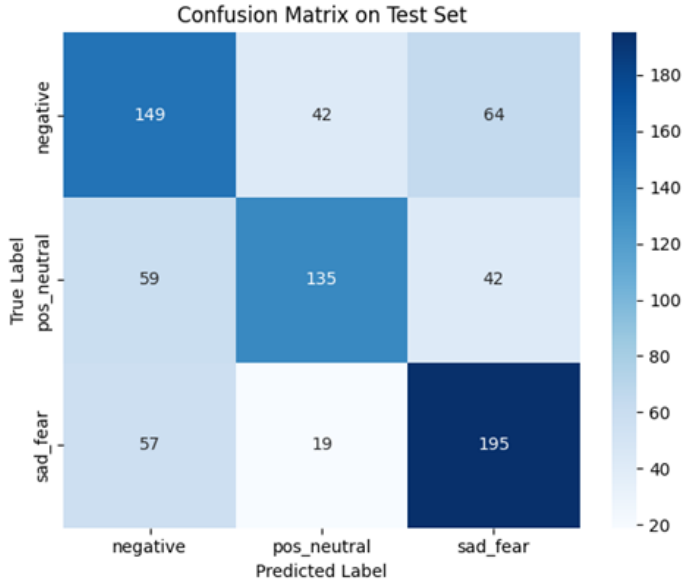

Fig. 4 Normalised confusion matrix for the three-class task.

Table 7. Cell values of the normalised confusion matrix (test set).

True \ Pred	Negative	Sad_Fear	Pos_Neutral
Negative	0.584	0.215	0.201
Sad_Fear	0.156	0.720	0.124
Pos_Neutral	0.221	0.207	0.572

5.5. Confusion-Matrix Examination

Comparisons are restricted to CNN-based SER with acoustic-only features, no fine-tuning of large transformers, and speaker-independent or cross-corpus evaluation (in Table 8). Representative figures: Satt et al. (RAVDESS, 6-class, 2-D CNN on log-Mel, Acc 0.624), Neumann & Vu [16] (IEMOCAP, 6-class, Att-CNN on log-Mel, speaker-leave-out, Acc 0.600), and Latif et al. (CREMA-D, 6-class, CNN+SE on MFCC, 5-fold speaker CV, Acc 0.610). Under a stricter multi-corpus three-class benchmark that merges CREMA-D, RAVDESS, and EmoDB with speaker-independent splits, the ResAttn1D-CNN on frozen wav2vec 2.0 attains 0.629 accuracy.

The improved outcome aligns with ablation evidence: a deeper residual hierarchy enlarges temporal receptive fields; SE channel attention re-weights affect-salient dimensions within the 768-D embedding space; and the Tri-Aug pipeline regularises against noise, temporal jitter, and masked spans—while maintaining a lightweight footprint (~3.9 M parameters) suitable for real-time inference.

Table 8. Reported accuracies of comparable CNN-based SER systems

Study	Dataset & Task	Model / Feature	Eval Protocol	Reported Acc
Satt et al. 2017	RAVDESS, 6-class	2-D CNN, log-Mel	Utterance-split	0.624
Neumann & Vu 2017	IEMOCAP, 6-class	Att-CNN, log-Mel	Speaker-leave-out	0.600
Latif et al. 2020	CREMA-D, 6-class	CNN + SE, MFCC	5-fold speaker CV	0.610
This work	CREMA-D + RAVDESS + EmoDB, 3-class	ResAttn1D-CNN, W2V2	Speaker-independent	0.629

6. Experimental Results and Analysis

6.1. Practical Implications

The configuration meets edge constraints while exceeding the 60% barrier under a speaker-independent, cross-corpus protocol. With ≈3.9 M parameters and frozen wav2vec 2.0 embeddings (768×400), inference is real-time (<3.5 ms/utterance, <15 MB GPU) and achieves 62.9% accuracy with 0.627 Macro-F1. Fixing inputs to 400 frames (central truncation/zero-padding) yields constant-latency serving. Training completes in ≈10 h with mixed precision and early stopping (~epoch 60) on laptop-class hardware. Freezing the encoder reduces memory and eliminates gradient flow through the feature stage; residual depth enlarges temporal context, SE attention re-weights affect-salient channels, and the lightweight Tri-Aug pipeline improves generalisation at negligible runtime cost. MCULP plus released manifests/scripts/Docker renders the benchmark deterministically reproducible and portable for swapping alternative backbones or features.

6.2. Current Limitations

Four main limitations of the present study are observed. First, class confusability remains, most notably between negative and pos_neutral; the row-normalised confusion matrix shows Negative→Pos_Neutral ≈ 0.201, with recalls Negative 0.584, Sad_Fear 0.720, Pos_Neutral 0.572. Second, the encoder is frozen (wav2vec 2.0-base), so the representation is not adapted to the emotion domain; this isolates downstream effects but limits potential gains from task-specific fine-tuning. Third, data scale and language coverage are modest: even after merging corpora, training remains at the ~6k-utterance level and languages are confined to English/German, leaving cross-lingual generalisation untested. Fourth, the system is acoustic-only; no multimodal cues (e.g., facial) or explicit utterance-level prosody descriptors are incorporated, which likely contributes to the residual negative ↔ pos_neutral ambiguity. These constraints define the scope of the reported results and inform the extensions outlined in the next subsection.

6.3. Acknowledge Current Limitations

Future work will explore lightweight adaptation of the frozen encoder—e.g., layer-wise discriminative rates or small adapters/projections for wav2vec 2.0 (and HuBERT)—while keeping the downstream CNN unchanged to capture domain-specific cues without increasing the footprint. Prosody-aware utterance features (pitch slope, energy contour, speaking rate, simple formant statistics) will be fused via a compact MLP to reduce the persistent negative ↔ pos_neutral ambiguity. Semi-supervised training on large unlabelled speech with confidence-filtered pseudo-labels and consistency regularisation across Tri-Aug views will expand data without manual annotation. Cross-lingual and cross-domain evaluation—including tonal/low-resource languages and noisier conditions—will quantify transfer and adaptation needs under the same protocol. On video-bearing datasets, lightweight late fusion of acoustic features with facial cues will be examined to strengthen cases where acoustic evidence is weak, while preserving real-time latency and low memory.

7. Conclusion

This study demonstrates that high-quality speech-emotion recognition (SER) can be achieved without resorting to heavyweight transformer fine-tuning. By introducing a Multi-Corpus Unified Labelling Protocol (MCULP), we harmonised CREMA-D, RAVDESS, and EmoDB into a balanced three-class benchmark, thereby eliminating label inconsistencies that have historically hindered cross-corpus evaluation. On top of frozen Wav2Vec 2.0 embeddings, we designed a five-block Residual Squeeze-and-Excitation 1-D CNN that integrates deep temporal context with channel-wise attention while remaining lightweight (3.9 M parameters, < 15 MB GPU memory). A composite Tri-Aug pipeline combining Gaussian noise, temporal cropping, and SpecAugment-style masking further improved generalisation under a strict speaker-independent split. Comprehensive experiments show

that the proposed system delivers 64.6 % validation accuracy, 62.9 % test accuracy, and a 0.627 macro-F1, outperforming a strong three-layer CNN baseline by > 20 percentage points and exceeding previously reported CNN results on comparable tasks. Ablation studies reveal that SE attention (+4.9 pp), Tri-Aug (+2.9 pp), and residual depth (+3.2 pp) contribute complementary gains, while learning-curve analysis confirms stable convergence within 10 h on a laptop-class GPU. Confusion-matrix inspection highlights residual ambiguity between low-arousal negative and neutral speech, suggesting the value of prosody-specific features and multimodal cues. All code, manifests, and Docker files will be released, enabling reproducible research and facilitating future extensions such as prosody-aware fusion, semi-supervised data expansion, and lightweight multimodal SER for edge deployment.

References

[1] Cao H, et al. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. IEEE Transactions on Affective Computing, 2014, 5 (4): 377–390.

[2] Livingstone S R, Russo F A. The RAVDESS: A Dynamic, Multimodal Set of Facial and Vocal Expressions. PLOS ONE, 2018, 13 (5): e0196391.

[3] Burkhardt F, et al. A Database of German Emotional Speech. In: Proc. Interspeech, 2005: 1517–1520.

[4] Baevski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems, 2020, 33: 12449–12460.

[5] Pepino F, Riera P, Ferrer C A. Emotion recognition from speech using Wav2Vec 2.0 embeddings. In: Proc. Interspeech, 2021: 3400–3404.

[6] Young S, Evermann G, Gales M, et al. The HTK Book. Cambridge: Cambridge University Engineering Department, 2006.

[7] Davis S B, Mermelstein P. Comparison of the parametric representation of speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28 (4): 357–366.

[8] Park D S, Chan W, Zhang Y, et al. SpecAugment: A simple data augmentation method for automatic speech recognition. In: Proc. Interspeech, 2019: 2613–2617.

[9] Satt A, Rozenberg S, Hoory R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In: Proc. Interspeech, 2017.

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[11] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[12] Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms. In: Proc. Interspeech, 2017.

[13] Loshchilov I, Hutter F. Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR), 2019.

[14] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[15] Scherer K R. Vocal communication of emotion: A review of research paradigms. Speech Communication, 2003, 40: 227–256.

[16] Neumann M, Vu N T. Attentive CNN-based Speech Emotion Recognition: A Study on IEMOCAP. In: Proc. Interspeech, 2017.