ISSN 2085-4552

Speech Emotion Recognition through Acoustic Data Augmentation and Attention-Driven CRNN-BiGRU

Muhamad Fitra Kacamarga[1], Kresna Andika Aprianto[2]
[1,2] Computer Science Department, BINUS Online Learning, Jakarta, Indonesia 11480
fitra.kacamarga@binus.ac.id[1], kresna.aprianto@binus.ac.id[2]

Accepted 21 June 2026
Approved 28 June 2026

***Abstract*— Speech emotion recognition (SER) systems have transformed human-computer interactions by enabling machines to identify emotional cues in speech. This study presents a *systematic* approach that combines complementary data-augmentation techniques with an efficient neural architecture to address data scarcity and model-robustness limitations. The proposed methodology applies a multi-dimensional augmentation framework that retains the original recordings while adding mildly perturbed copies targeting acoustic-environment variability (background-noise injection), speaking-rate variability (time stretching at rates of 0.9 and 1.1), and vocal-pitch variability (pitch shifting by one to two semitones). This augmented dataset feeds into a novel Convolutional Recurrent Neural Network (CRNN) integrated with a Bidirectional Gated Recurrent Unit (BiGRU) and a Bahdanau attention mechanism, designed to capture both local spectral and temporal emotional features effectively. The lightweight model (166,729 parameters, ≈0.64 MB) processes log-Mel spectrograms with Voice Activity Detection. On the RAVDESS database, the model achieves a weighted accuracy (WA) of 85.42% and an unweighted accuracy (UA) of 85.35%; data augmentation provides a consistent +1.74-point WA improvement over the non-augmented baseline, and the model outperforms comparable prior methods (e.g., +6.1 points WA over a CNN with Multi-Head attention) while using far fewer parameters than typical deep SER models. These results validate the effectiveness of integrating systematic data augmentation with an efficient neural architecture for SER applications.**

Index Terms*— *Speech Emotion Recognition; CRNN; BiGRU; Attention Mechanism; Data Augmentation; Deep Learning; Audio Processing; Log-Mel Spectrogram.

I. INTRODUCTION

Human-Computer Interaction (HCI) encompasses the study, planning, and design of the interaction between users and computers. While traditional HCI focuses on efficiency and usability, modern systems increasingly aim to emulate the nuanced characteristics of human-to-human communication, particularly in recognizing and responding to emotional states. Among the various modalities of HCI, three key domains have emerged as critical for natural interactions: speech processing [1]–[7], textual analysis [8], [9], and facial recognition [10], [11]. The effective integration of these modalities enables systems to comprehend and respond to diverse human experiences [12].

Speech Emotion Recognition (SER), an automated process for identifying emotional states from voice signals, has emerged as a promising advancement in this field. Recent developments in deep learning and signal processing have created new opportunities for SER applications, including healthcare and wellness for psychological assessment [14], consumer technology for smart-home automation [13], and professional settings for human-robot interaction [15]. This broad spectrum demonstrates the transformative potential of SER in enhancing human-computer interactions.

Despite these advances, current SER systems face significant limitations. Unlike Automatic Speech Recognition (ASR), which benefits from vast standardized datasets, SER research is impeded by data scarcity due to the expense of recording emotions in controlled environments and the prevalence of noisy real-life data. The subjective nature of emotions, influenced by language and culture, further increases annotation complexity [21]. Recognizing emotions in spontaneously expressed speech also poses a greater challenge than scripted performances [22], [23], particularly under environmental noise and cross-cultural variation.

A typical SER system comprises three components: a comprehensive corpus of speech emotions, effective feature-extraction techniques, and efficient classification algorithms. Commonly employed features include amplitude, Mel-frequency cepstral coefficients (MFCC), log-Mel spectrograms, and the zero-crossing rate [24]. While traditional methods such as Support Vector Machines and Hidden Markov Models have been used for acoustic-feature classification, the advent of deep learning has driven substantial progress. For example, Lee *et al.* [25] achieved 63.89% accuracy using Bidirectional Long Short-Term Memory (BiLSTM) networks, while Lim *et al.* [26] reached 86.06% on the EMO-DB database with convolutional neural networks.

To address these persistent challenges, this study introduces two key innovations:

A systematic data augmentation framework that enhances model robustness through strategic, multi-dimensional manipulation of acoustic properties while retaining the original samples.

A lightweight CRNN-BiGRU architecture (166,729 parameters) that integrates local feature extraction with temporal pattern recognition and an attention mechanism.

These innovations address significant deficiencies in contemporary SER research, namely the need for reliable performance across varied acoustic environments and an efficient model design suitable for practical deployment.

II. Literature Review

Speech Emotion Recognition (SER) is the task of automatically identifying human emotions from spoken utterances. Over the past five years, deep-learning approaches have significantly improved SER performance [1]. Models based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can learn directly from raw audio or spectrograms, eliminating manual feature engineering [2].

2.1 CNN-Based Models

CNNs excel at extracting local spectral-temporal patterns through learned filters. Bhangale and Kothandaraman (2023) developed a lightweight 1-D CNN that achieved 93–94% accuracy on the Emo-DB and RAVDESS datasets [4]. The primary advantages of CNN-based SER include automatic feature learning and a degree of noise robustness through pooling operations [5]. However, CNNs with a limited receptive field may not capture the long-range temporal dependencies that are critical for emotion recognition, as emotions can evolve over an entire utterance.

2.2 RNN-Based Models

RNNs directly model the temporal sequence of acoustic features, making them well suited to capturing the dynamic patterns of emotion in speech. Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) are the most prominent RNN variants in recent SER research. Trinh *et al.* (2022) compared a CNN, a CNN-RNN hybrid, and a pure GRU model on the IEMOCAP corpus; the GRU network achieved the highest accuracy (97.47%), outperforming both the CNN and the CNN+RNN models under identical conditions [7], demonstrating that the GRU gating mechanism effectively captures emotional dynamics over time.

2.3 Hybrid CNN-RNN (CRNN) Models

Hybrid architectures that combine convolutional and recurrent layers have become a cornerstone of modern SER systems. CNN layers first extract high-level features from spectrogram segments, and RNN layers then process the sequence of these features over time, leveraging both local feature extraction and global temporal modeling [8], [9]. Yadav *et al.* (2021) proposed a convolutional recurrent network using CNN followed by a bidirectional LSTM; with noise-reduction preprocessing and carefully chosen augmentation, their CRNN attained 85.8% accuracy on RAVDESS and 91.6% on Emo-DB [10], [11]. Another study reported approximately 90% accuracy on RAVDESS [12].

2.4 Bidirectional RNNs and BiGRU

Bidirectional architectures (such as BiLSTM or BiGRU) process sequences in both forward and backward directions, providing access to past and future context simultaneously. Xu *et al.* (2024) presented an SER model employing a Bi-GRU with a multi-head attention mechanism that significantly outperformed comparable unidirectional models across multiple metrics [14]–[16], emphasizing the benefit of capturing the complete temporal context of an utterance.

2.5 Attention Mechanisms

Attention mechanisms allow networks to focus on the most salient parts of a speech signal. In an RNN-based SER model, an attention layer learns to weigh the importance of each time frame before the final emotion prediction — valuable in speech, where not all moments are equally informative [26]. In addition to RNNs, pure attention-based architectures have emerged: a recent parallel CNN-transformer architecture achieved 89.3% accuracy, outperforming a comparable CNN–BiLSTM with attention under identical conditions [23].

2.6 Data Augmentation Techniques

Data augmentation addresses the limited size and imbalance of emotion datasets and improves model robustness. Common methods include:

Noise Injection: Adding background or white Gaussian noise to simulate different acoustic environments [24].

Pitch and Speed Variations: Altering pitch or playback speed to mimic different speakers or speaking styles [25].

Room and Channel Effects: Applying filters to simulate different recording conditions.

Segment Augmentation: Splitting utterances or random cropping to increase the sample count.

Empirical evidence shows that augmentation can substantially boost SER performance. Yadav *et al.* (2021) improved their CNN-BiLSTM model from the low 80s to 85.8% on RAVDESS [26], and an augmented CNN model achieved 92.6% on RAVDESS [27]

III. METHOD

This study adopts an experimental methodology to assess a Convolutional Recurrent Neural Network (CRNN) integrated with a Bidirectional Gated Recurrent Unit (BiGRU) for the classification of a speech-emotion dataset. The stages of the study — data preparation, data augmentation, feature extraction, model training, and evaluation — are depicted in Fig. 1. The model's performance is evaluated on the test set using weighted and unweighted accuracy, and is further analyzed through a confusion matrix.

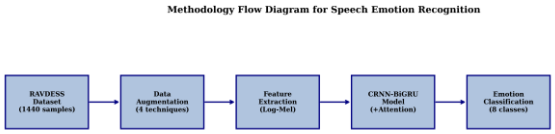

Fig. 1. *Flow diagram of the proposed speech-emotion-recognition methodology, showing the pipeline from dataset input through augmentation, feature extraction, model training, and emotion classification.*

3.1 Emotion Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was used in this study. Developed at Ryerson University in Canada [30], the full database is a high-quality collection designed for emotion-recognition research. This work uses the **speech audio** subset, which comprises **1,440 recordings** from 24 professional actors (12 male and 12 female), all native North American English speakers.

TABLE I. **RAVDESS (SPEECH SUBSET) STATISTICS**

Attribute	Value
Total audio files	1,440
Number of emotion classes	8
Number of actors	24 (12 male, 12 female)
Files per actor	60
Sampling rate	48 kHz
Recording environment	Professional soundproof booth

TABLE II. **EMOTION CLASS DISTRIBUTION**

Emotion	Code	Sample Count	Percentage
Neutral	1	96	6.67%
Calm	2	192	13.33%
Happy	3	192	13.33%
Sad	4	192	13.33%
Angry	5	192	13.33%
Fearful	6	192	13.33%
Disgust	7	192	13.33%
Surprised	8	192	13.33%

Dataset balance analysis. The RAVDESS speech subset exhibits a near-balanced distribution across emotion classes. The neutral class contains 96 samples (normal intensity only), whereas the other seven classes contain 192 samples each (96 normal + 96 strong intensity). This slight imbalance (a neutral-to-other ratio of 1:2) is inherent to the dataset design, since neutral expressions lack the intensity variation present in the other emotions. The dataset nonetheless maintains balanced representation across speakers (60 files per actor) and gender (equal male/female distribution), which is why no class-weighting strategy was required.

The dataset's key characteristics for SER research are: (1) **emotional diversity** — eight expressions (neutral, calm, happy, sad, angry, fearful, disgust, surprised); (2) **controlled conditions** — all recordings captured in a soundproof booth at 48 kHz; (3) **standardized lexical content** — each actor utters the same two statements, *"Kids are talking by the door"* and *"Dogs are sitting by the door,"* preventing lexical bias; and (4) **two intensity levels** (normal and strong) for all emotions except neutral.

3.2 Systematic Data Augmentation Framework

Prior to feature extraction, a systematic multi-dimensional augmentation framework was applied to the training data. Here, *"systematic"* denotes a framework that targets several distinct acoustic dimensions of speech variability rather than a single transformation. Crucially, the framework **retains the original (clean) recordings and adds mildly perturbed copies** (two augmented copies per training utterance); the model therefore learns from both the clean distribution and its variations, which avoids the train/test distribution mismatch that overly aggressive augmentation can cause. Each augmented copy applies the following transforms probabilistically, illustrated in Fig. 2:

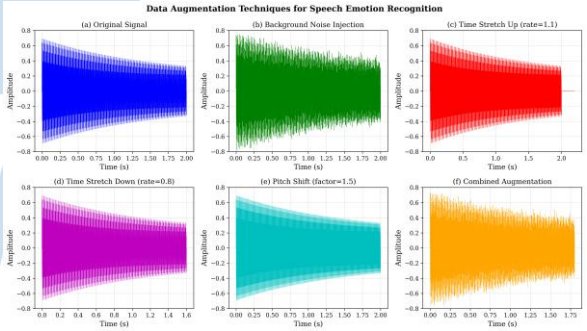

Fig. 2. *Illustration of the data-augmentation techniques applied to a speech sample: (a) original signal, (b) background-noise injection, (c)–(d) time stretching, (e) pitch shifting, and (f) combined augmentation.*

Acoustic-environment variability (background-noise injection): low-level Gaussian noise (amplitude factor 0.02 of the signal peak) added with 30% probability per copy, to simulate diverse recording environments.

Speaking-rate variability (time stretching): the time axis is stretched with 50% probability at a rate of either 0.9 or 1.1, to account for slower or faster speaking patterns.

Vocal-pitch variability (pitch shifting): the pitch is shifted with 50% probability by ±1–2 semitones, to simulate speaker differences.

This multi-dimensional strategy helps the model learn emotional features that are robust to recording

conditions and speaker variation; retaining the clean samples ensures the training distribution remains aligned with the (clean) test distribution.

3.3 Feature Extraction

3.3.1 Feature Selection: Log-Mel Spectrogram vs. MFCC

This study employs log-Mel spectrograms rather than Mel-Frequency Cepstral Coefficients (MFCCs) as the primary acoustic feature, for the following technical reasons:

Preservation of spectral structure. Log-Mel spectrograms retain the complete time–frequency representation of speech, including harmonic relationships and spectral dynamics that are important for emotion recognition. MFCCs apply a Discrete Cosine Transform (DCT) that decorrelates and compresses this information, potentially discarding fine-grained spectral patterns that encode emotional state.

Compatibility with deep learning. CNNs are designed to learn hierarchical features from structured input. Log-Mel spectrograms preserve the spatial relationship between frequency bands, allowing convolutional filters to capture local spectral–temporal patterns effectively.

Retention of prosodic information. The Mel filterbank preserves perceptually relevant frequency resolution along with fundamental frequency (F0), formant transitions, and energy contours — critical emotional indicators. The cepstral analysis in MFCCs tends to remove pitch-related information that helps distinguish, for example, anger (high F0) from sadness (low F0).

State-of-the-art practice. Recent high-performing SER systems consistently employ log-Mel or raw spectrograms as input features [4], [32].

TABLE III. **FEATURE-EXTRACTION PARAMETERS**

Parameter	Value
Number of Mel bands	64
Maximum frequency	6,000 Hz
Window length	2,048 samples
Hop length	1,365 samples (2048 × 2/3)
FFT size	4,096

The extraction process converts raw audio waveforms into time-domain signals and then applies Voice Activity Detection (VAD) to identify the significant portions of the signal. The result is a truncated log-Mel spectrogram that focuses on the most pertinent segments. A representative waveform and its corresponding log-Mel spectrogram are shown in Fig. 3.

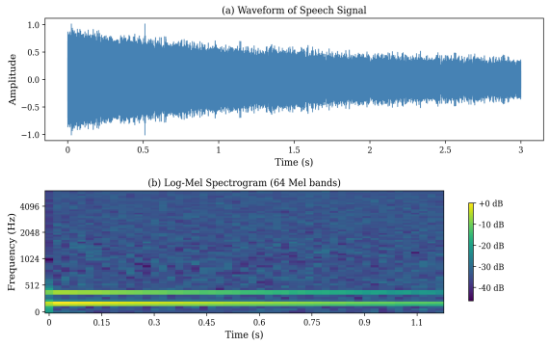

Fig. 3. *(a) Waveform of a speech sample after preprocessing and (b) the corresponding log-Mel spectrogram (64 Mel bands) used as the model input.*

3.4 Model Design

The proposed model combines the strengths of a CRNN for local feature extraction with a BiGRU and an attention mechanism for global temporal modeling. The complete architecture is shown in Fig. 4.

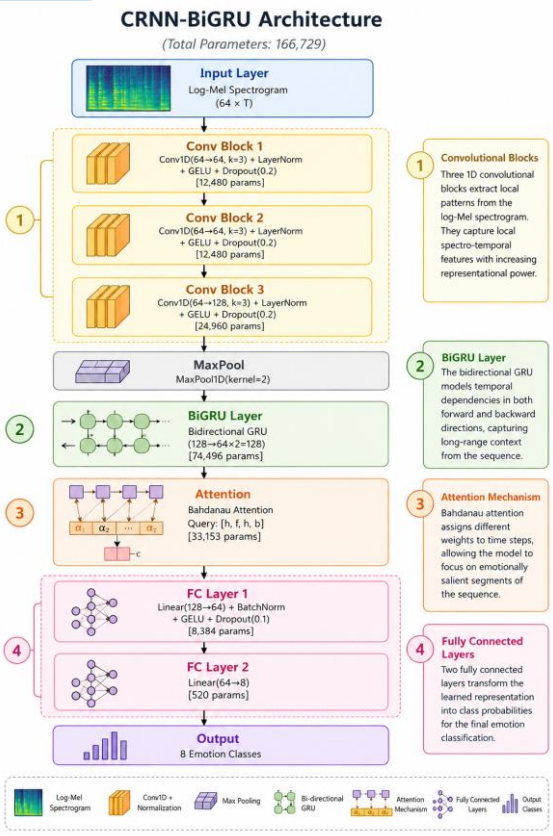

Fig. 4. *Detailed architecture of the proposed CRNN-BiGRU model, showing four main components: (1) three convolutional blocks for local feature extraction from the log-Mel spectrogram, (2) a bidirectional GRU layer for temporal modeling, (3) a Bahdanau attention mechanism for focusing on emotionally salient segments, and (4) fully connected layers for classification. Total parameters: 166,729.*

3.4.1 Architecture Justification

CRNN vs. CNN or RNN alone. The hybrid CRNN architecture was chosen on the basis of evidence that CNNs excel at extracting local spectral features while RNNs capture temporal dependencies. CNN-only approaches [34] (71.6% accuracy) do not adequately model temporal relationships in emotional speech, whereas RNN-only methods [33] (69.4% UA) struggle to extract robust local features. The CRNN combination leverages the complementary strengths of both.

BiGRU vs. LSTM/BiLSTM. BiGRU was chosen for three reasons: (i) *computational efficiency* — GRUs have fewer parameters than LSTMs, reducing training time while maintaining comparable performance; (ii) *gradient stability* — GRU gating mechanisms preserve long-term dependencies; and (iii) *bidirectional processing* — capturing both past and future context, which is important where anticipatory and carryover emotional cues span the utterance.

Bahdanau attention vs. other mechanisms. Bahdanau attention was selected after experimental evaluation, offering dynamic focus on emotionally salient segments, the ability to emphasize relevant acoustic features while down-weighting less informative portions, and stable gradient flow during training. The attention mechanism addresses a limitation of standard RNN approaches [25], which use only the final hidden state for classification and may lose emotional information distributed throughout the utterance.

ConvLN and GELU activation. Layer Normalization was chosen over Batch Normalization for stability with the variable-length sequences common in speech, and the GELU activation was selected for its smoother gradient properties, which improved performance in preliminary testing relative to ReLU.

3.4.2 Architecture Details

The architecture comprises:

Three convolutional blocks, each containing Conv1D + Layer Normalization + GELU + Dropout (0.2):

ConvBlock 1: 64 → 64 channels, kernel = 3

ConvBlock 2: 64 → 64 channels, kernel = 3

ConvBlock 3: 64 → 128 channels, kernel = 3

MaxPool1D (kernel = 2) for temporal downsampling.

Bidirectional GRU (input 128, hidden 64 per direction → 128-dimensional output) preceded by Layer Normalization.

Bahdanau attention, which uses the concatenated final BiGRU hidden states as the query over all BiGRU outputs, with attention weights obtained via a softmax.

Classification head: Linear(128 → 64) + BatchNorm + GELU + Dropout(0.1), followed by Linear(64 → 8 classes).

TABLE IV. **PROPOSED-MODEL COMPLEXITY ANALYSIS**

Component	Parameters	Percentage
Convolutional blocks (3)	49,920	29.9%
Bidirectional GRU	74,496	44.7%
Bahdanau attention	33,153	19.9%
Fully connected layers	8,904	5.3%
Layer normalization	256	0.2%
Total	**166,729**	**100%**
Model size (FP32)	**≈0.64 MB**	—

With only **166,729 trainable parameters (≈0.64 MB)**, the proposed model is markedly more lightweight than typical deep SER systems, which commonly rely on millions of parameters (e.g., deep 2-D CNNs or transformer/self-supervised speech models). This compact design enables deployment on resource-constrained devices while preserving competitive accuracy, and constitutes the model-complexity contribution of this work.

3.5 Model Training

The reported results were obtained using a **stratified hold-out** evaluation: the RAVDESS speech subset was partitioned into a training set (80%) and a test set (20%) at the utterance level, stratified by emotion to preserve the class proportions, with a fixed random seed for reproducibility. All data were shuffled prior to partitioning. Data augmentation (Section 3.2) was applied to the **training partition only**; the test partition consists of clean recordings.

The training configuration was held constant across experiments:

Optimizer: AdaBelief (learning rate 1×10^{-3}, β = (0.9, 0.999), weight decay 1×10^{-7})

Loss: label-smoothing cross-entropy (smoothing = 0.1)

Learning-rate scheduler: ReduceLROnPlateau (factor 0.7, patience 2, cooldown 3)

Epochs: 100; mini-batch size: 32

We note that the present evaluation uses an utterance-level split in which the 24 actors appear in both the training and test partitions; a fully speaker-independent evaluation (e.g., leave-speaker-group-out cross-validation) is identified as an important direction for future work.

3.6 Model Evaluation

After training, the model was evaluated on the held-out test set using the following metrics.

Weighted Accuracy (WA) — the proportion of correctly classified samples, weighted by class proportion:

$$WA = \frac{\sum_{i=1}^{C} n_i \cdot r_i}{\sum_{i=1}^{C} n_i}$$

where n_i is the number of samples in class i, r_i is the recall for class i, and C is the number of classes.

Unweighted Accuracy (UA) — the average per-class accuracy, giving equal importance to each class regardless of its frequency:

$$UA = \frac{1}{C}\sum_{i=1}^{C} r_i = \frac{1}{C}\sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i}$$

where TP_i and FN_i are the true positives and false negatives for class i. A confusion matrix was also generated to visualize the distribution of true and predicted labels across the eight emotion classes.

IV. Result and Discussions

The model was implemented in PyTorch and trained for 100 epochs with a batch size of 32. Table V reports the test performance with and without data augmentation.

TABLE V. **Results With and Without Data Augmentation**

Data	WA (%)	UA (%)
Non-Augmented	83.68	83.78
Augmented	85.42	85.35

Data augmentation improves performance by **+1.74 percentage points in WA** (83.68% → 85.42%) and +1.57 points in UA (83.78% → 85.35%). The gain is most pronounced for the harder classes — for example, the recall for *fearful* rises from 79.5% to 92.3% and for *calm* from 87.2% to 92.3% — indicating that augmentation broadens the effective coverage of the training distribution and encourages features that are less dependent on the original recording conditions. Retaining the clean samples alongside the augmented copies is important: it keeps the training distribution aligned with the clean test set, so that augmentation improves rather than degrades generalization.

Confusion-matrix analysis. The model (with augmentation) shows strong, well-balanced performance across the eight RAVDESS emotion classes, with pronounced diagonal dominance (Fig. 5):

Highest accuracy: disgust (97.4%), calm (92.3%), fearful (92.3%), and angry (92.1%).

Strong performance: surprised (86.8%) and neutral (84.2%).

Hardest classes: happy (69.2%) and sad (68.4%).

The most common confusions occur between acoustically similar emotion pairs: *happy* is most often mistaken for *surprised* (15.4%) — both high-arousal expressions — while *sad* is confused with *fearful* (15.8%) and *disgust* (7.9%). These results highlight *happy* and *sad* as the priority targets for future feature engineering.

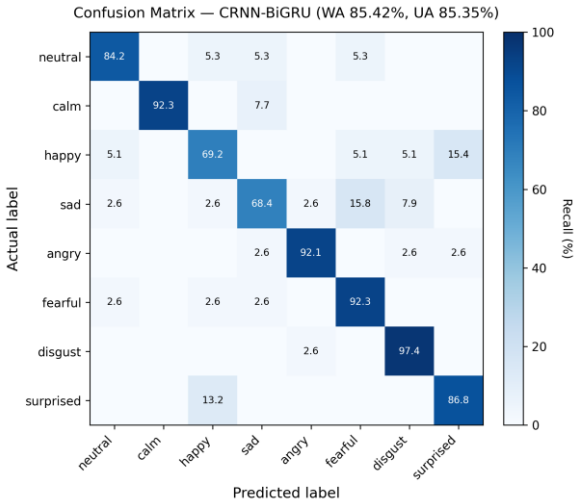

Fig. 5. *Confusion matrix (row-normalized recall, %) of the proposed CRNN-BiGRU model with data augmentation on the RAVDESS test set (WA 85.42%, UA 85.35%). Rows are the actual labels and columns are the predicted labels.*

Comparative performance. Table VI compares the proposed model with previously reported results on the RAVDESS dataset.

TABLE VI. **Comparison of Accuracy With Other Studies on RAVDESS**

Method	WA (%)	UA (%)	Year
GResNet + Multi-Task [28]	64.5	64.6	2019
BLSTM + Capsule Routing [33]	—	69.4	2019
Multimodal Fine-grained Learning [29]	77.0	74.7	2020
DCGAN + Augmentation [31]	72.3	72.3	2023
Head Fusion [30]	77.8	77.4	2023
CNN + Multi-Head [32]	79.3	78.2	2023
CRNN-BiGRU (proposed)	**83.68**	**83.78**	2025
CRNN-BiGRU + Augmentation (proposed)	**85.42**	**85.35**	2025

Relative to these methods, the proposed CRNN-BiGRU with augmentation is competitive while being far more compact — for example, +6.1 percentage points in WA over CNN + Multi-Head [32] (79.3%) and +7.6 points over Head Fusion [30] (77.8%). The gains are consistent with the architecture's design: the CRNN integrates spectral and temporal information, the BiGRU captures the temporal evolution of emotions, and the attention mechanism concentrates on emotionally salient frames. Crucially, these results are achieved with only 166,729 parameters (≈0.64 MB), making the proposed model an order of magnitude more compact than the deep CNN and attention-based systems it is compared against; this favorable

accuracy-to-size trade-off is the central contribution of this work

V. CONCLUSIONS

This research proposes a structure that combines Convolutional Neural Networks with a Bidirectional Gated Recurrent Unit, a Bahdanau attention mechanism, and a systematic data-augmentation framework for speech emotion recognition on the RAVDESS dataset. The incorporation of attention enhances the model's focus on key speech segments, while the lightweight design (166,729 parameters, ≈0.64 MB) keeps the computational cost low. Evaluated by weighted and unweighted accuracy, the model achieves a WA of 85.42% and a UA of 85.35%, outperforming the comparable prior approaches summarized in Table VI (all ≤ 79.3% WA) while using far fewer parameters, with data augmentation contributing a +1.74-point WA gain over the non-augmented baseline.

The principal contributions are: (1) a systematic augmentation framework that targets multiple acoustic dimensions while retaining the original samples, yielding a consistent accuracy gain; (2) an efficient CRNN-BiGRU architecture with Bahdanau attention suitable for resource-constrained deployment; and (3) a detailed model-complexity analysis demonstrating that a carefully designed CNN-RNN architecture can reach competitive accuracy at a small fraction of the parameter budget of typical deep SER models.

The model nonetheless has limitations. It can struggle with acoustically similar expressions (e.g., happy vs. surprised and sad vs. fearful), and the present evaluation uses an utterance-level stratified hold-out in which speakers are shared between the training and test partitions; a fully speaker-independent assessment is left for future work. Future research could also explore augmentation tailored to specific emotion categories and multimodal approaches that combine acoustic features with complementary information sources.

REFERENCES

[1] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, ISCA, Sep. 2014, pp. 223–227, doi: 10.21437/Interspeech.2014-57.

[2] M. Chen, X. He, J. Yang, and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246.

[3] X. Wu *et al.*, "Speech Emotion Recognition Using Capsule Networks," in *ICASSP 2019*, IEEE, May 2019, pp. 6695–6699, doi: 10.1109/ICASSP.2019.8683163.

[4] Y. Xu, H. Xu, and J. Zou, "HGFM: A Hierarchical Grained and Feature Model for Acoustic Emotion Recognition," in *ICASSP 2020*, IEEE, May 2020, pp. 6499–6503, doi: 10.1109/ICASSP40776.2020.9053039.

[5] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention Driven Fusion for Multi-Modal Emotion Recognition," in *ICASSP 2020*, IEEE, May 2020, pp. 3227–3231, doi: 10.1109/ICASSP40776.2020.9054441.

[6] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-Head Attention for Speech Emotion Recognition with Auxiliary Learning of Gender Recognition," in *ICASSP 2020*, IEEE, May 2020, pp. 7179–7183, doi: 10.1109/ICASSP40776.2020.9054073.

[7] M. Xu, F. Zhang, and S. U. Khan, "Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, Jan. 2020, pp. 1058–1064, doi: 10.1109/CCWC47524.2020.9031207.

[8] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding Emotions in Text Using Deep Learning and Big Data," *Comput. Human Behav.*, vol. 93, pp. 309–317, Apr. 2019, doi: 10.1016/j.chb.2018.12.029.

[9] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-Emotion Neural Network for Emotion Recognition From Text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019, doi: 10.1109/ACCESS.2019.2934529.

[10] J. Yang, F. Zhang, B. Chen, and S. U. Khan, "Facial Expression Recognition Based on Facial Action Unit," in *2019 Tenth International Green and Sustainable Computing Conference (IGSC)*, IEEE, Oct. 2019, pp. 1–6, doi: 10.1109/IGSC48788.2019.8957163.

[11] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jul. 2017, pp. 522–531, doi: 10.1109/CVPRW.2017.79.

[12] H. L. O'Brien, I. Roll, A. Kampen, and N. Davoudi, "Rethinking (Dis)engagement in human-computer interaction," *Comput. Human Behav.*, vol. 128, p. 107109, Mar. 2022, doi: 10.1016/j.chb.2021.107109.

[13] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-Time Speech Emotion Analysis for Smart Home Assistants," *IEEE Trans. Consum. Electron.*, vol. 67, no. 1, pp. 68–76, Feb. 2021, doi: 10.1109/TCE.2021.3056421.

[14] L.-S. A. Low, M. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of Clinical Depression in Adolescents' Speech During Family Interactions," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586, Mar. 2011, doi: 10.1109/TBME.2010.2091640.

[15] X. Huahu, G. Jue, and Y. Jian, "Application of Speech Emotion Recognition in Intelligent Household Robot," in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, IEEE, Oct. 2010, pp. 537–541, doi: 10.1109/AICI.2010.118.

[16] W.-J. Yoon, Y.-H. Cho, and K.-S. Park, "A Study of Speech Emotion Recognition and Its Application to Mobile Services," in *Ubiquitous Intelligence and Computing*, Springer, pp. 758–766, doi: 10.1007/978-3-540-73549-6_74.

[17] S. G. Koolagudi, K. Sharma, and K. Sreenivasa Rao, "Speaker Recognition in Emotional Environment," 2012, pp. 117–124, doi: 10.1007/978-3-642-32112-2_15.

[18] P. Ekman and H. Oster, "Facial Expressions of Emotion," *Annu. Rev. Psychol.*, vol. 30, no. 1, pp. 527–554, Jan. 1979, doi: 10.1146/annurev.ps.30.020179.002523.

[19] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: 10.1016/j.patcog.2010.09.020.

[20] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, Sep. 1977, doi: 10.1016/0092-6566(77)90037-X.

[21] R. Altrov and H. Pajupuu, "The influence of language and culture on the understanding of vocal emotions," *Journal of*

Estonian and Finno-Ugric Linguistics, vol. 6, no. 3, pp. 11–48, Dec. 2015, doi: 10.12697/jeful.2015.6.3.01.

[22] Z. Zhao *et al.*, "Self-attention transfer networks for speech emotion recognition," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 1, pp. 43–54, Feb. 2021, doi: 10.1016/j.vrih.2020.12.002.

[23] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition," in *Interspeech 2018*, ISCA, Sep. 2018, pp. 3087–3091, doi: 10.21437/Interspeech.2018-1242.

[24] A. Sandesara, S. Parikh, P. Sapovadiya, and M. Rahevar, "A Comparative Study On Speech Emotion Recognition," *Int. J. Res. Eng. Sci. Manag.*, vol. 3, no. 11, pp. 25–35, Nov. 2020, doi: 10.47607/ijresm.2020.366.

[25] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*, ISCA, Sep. 2015, pp. 1537–1540, doi: 10.21437/Interspeech.2015-336.

[26] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, IEEE, Dec. 2016, pp. 1–4, doi: 10.1109/APSIPA.2016.7820699.

[27] X. Han, F. Chen, and J. Ban, "Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure," *Electronics*, vol. 12, no. 4, p. 978, Feb. 2023, doi: 10.3390/electronics12040978.

[28] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimed. Tools Appl.*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019, doi: 10.1007/s11042-017-5539-3.

[29] H. Li, W. Ding, Z. Wu, and Z. Liu, "Learning Fine-Grained Cross Modality Excitement for Speech Emotion Recognition," Oct. 2020.

[30] M. Xu, F. Zhang, and W. Zhang, "Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021, doi: 10.1109/ACCESS.2021.3067460.

[31] J.-Y. Baek and S.-P. Lee, "Enhanced Speech Emotion Recognition Using DCGAN-Based Data Augmentation," *Electronics*, vol. 12, no. 18, p. 3966, Sep. 2023, doi: 10.3390/electronics12183966.

[32] R. Ullah *et al.*, "Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer," *Sensors*, vol. 23, no. 13, p. 6212, Jul. 2023, doi: 10.3390/s23136212.

[33] Md. A. Jalal, E. Loweimi, R. K. Moore, and T. Hain, "Learning Temporal Clusters Using Capsule Routing for Speech Emotion Recognition," in *Interspeech 2019*, ISCA, Sep. 2019, pp. 1701–1705, doi: 10.21437/Interspeech.2019-3068.

[34] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, p. 101894, May 2020, doi: 10.1016/j.bspc.2020.101894