Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084015

ARTICLE

DDE-SER: A Dual-Decomposition Ensemble Framework Fusing Adaptive Variational Modes and Harmonic-Percussive Spectrograms for Speech Emotion Recognition

David Hason Rudd[1,*], **Cesar Sanin**[2], **Md Rafiqul Islam**[3], **Xianzhi Wang**[1] and **Huan Huo**[1]

[1]School of Computer Science, The University of Technology Sydney, 15 Broadway, Ultimo, NSW, Australia
[2]School of Science and Technology, University of New England, Elm Avenue, Armidale, NSW, Australia
[3]Faculty of Science and Technology, Charles Darwin University, 54 Cavenagh St., Darwin, NT, Australia

*Corresponding Author: David Hason Rudd. Email: david.hasonrudd@uts.edu.au

Received: 15 April 2026; Accepted: 18 June 2026; Published: 13 August 2026

ABSTRACT: The accurate classification of human emotions from speech remains a formidable challenge due to the dynamic, non-stationary properties of audio signals and pervasive background noise. Traditional single-domain extraction methods frequently fail to capture overlapping acoustic phenomena, resulting in high misclassification rates among acoustically similar emotions. To overcome this, we propose the Dual-Decomposition Ensemble (DDE-SER), an architecture that synergizes 1D adaptive frequency filtering with 2D spatial spectrogram separation. The framework operates through two distinct pipelines: an adaptive time-domain branch that leverages VGG-optiVMD to autonomously extract Intrinsic Mode Functions (IMFs), and a structural spectrogram branch that applies orthogonal median filtering to decouple continuous harmonic formants from transient percussive noises. The primary novel contribution is a trainable Gated Attention mechanism that dynamically fuses these two previously established orthogonal decomposition pipelines based on the underlying emotional context, mitigating feature redundancy without the parameter overhead of large foundation models. Additionally, an acoustic perturbation strategy involving targeted pitch shifts and noise injection is applied to prevent overfitting on pristine studio datasets. Validated using a rigorous Speaker-Independent Cross-Validation (SICV) methodology, DDE-SER demonstrates competitive predictive performance, achieving an overall accuracy of 85.30% on the EMO-DB corpus and 62.17% on the seven-class RAVDESS corpus. Diagnostic results confirm the framework partially mitigates ambiguities between high-arousal states, such as Anger and Happiness, while maintaining the lightweight computational efficiency required for affective computing deployments.

KEYWORDS: Acoustic emotion recognition; variational mode decomposition; harmonic-percussive separation; feature fusion; ensemble learning; convolutional neural networks

1 Introduction

Speech Emotion Recognition (SER) is a foundational component of affective computing, enabling more natural and empathetic Human-Computer Interaction (HCI) across robotics, psychiatric assessment, and automated service systems [1,2]. However, reliably decoding emotional states from voice signals is inherently difficult. Human speech is highly non-stationary, and the extraction of salient affective cues is frequently disrupted by inter-speaker variability and environmental noise [3,4].

Historically, SER methodologies have relied on static spectral features, such as Mel-Frequency Cepstral Coefficients (MFCCs) and standard Mel-spectrograms, processed through conventional classifiers or Convolutional Neural Networks (CNNs) [5,6]. While deep learning has improved predictive performance, treating the speech signal as a monolithic input presents significant limitations. Raw spectrograms overlay continuous tonal energy (harmonics) and transient noise (percussives) within the same spatial dimension. Processing this overlapping data forces neural networks to untangle conflicting acoustic information simultaneously, which severely degrades classification accuracy between high-arousal states, like Anger and Happiness, that share similar energy envelopes but possess fundamentally distinct micro-structures.

To bypass these limitations, contemporary research has shifted toward multi-domain signal decomposition prior to deep feature extraction. In the 1D time-domain, Variational Mode Decomposition (VMD) mathematically mitigates mode-mixing by isolating discrete, band-limited Intrinsic Mode Functions (IMFs) [7,8]. We previously demonstrated that automating this process via gradient feedback (VGG-optiVMD) optimally preserves emotional cues [9]. In the 2D spatial domain, Harmonic-Percussive (HP) separation isolates continuous pitch from explosive transients using orthogonal median filtering [10]. Yet, these two powerful decomposition techniques remain isolated in the literature, and no prior work has attempted to bridge them to partially mitigate class confusion without the computational overhead of large-scale foundation models [11,12].

To bridge this gap, this study proposes a mathematically structured ensemble approach. We introduce the Dual-Decomposition Ensemble (DDE-SER), which unites 1D adaptive frequency decomposition with 2D structural time-frequency mapping. The primary contributions of this research are:

1. Building upon our prior feature extraction methodologies [9,10], we propose a dual-branch SER architecture in which the primary novel contribution is the gated fusion mechanism that bridges the established VGG-optiVMD and HP-Mel pipelines. This orchestration captures orthogonal affective features simultaneously, with substantially fewer parameters than foundation-model approaches.
2. We introduce a learnable Gated Attention mechanism that dynamically prioritizes multidimensional features based on the latent emotional class, outperforming standard linear fusion.
3. We deploy targeted, in-place data augmentation (pitch shifting and noise injection) to enhance model generalization, avoiding the pitfalls of redundant blind oversampling.
4. Through rigorous empirical evaluation utilizing a strict Speaker-Independent Cross-Validation (SICV) protocol, we demonstrate that the DDE-SER framework achieves an overall accuracy of 85.30 ± 13.01% standard deviation (SD) on EMO-DB and 62.17 ± 13.93% (SD) on the seven-class RAVDESS corpus, partially mitigating persistent misclassifications between acoustically similar high-arousal states.

The remainder of this paper is organized as follows. Section 2 reviews related literature. Section 3 details the DDE-SER methodology and computational framework. Section 4 outlines the datasets and experimental setup. Section 5 presents the performance analysis, and Section 6 concludes the study.

2 Related Work

SER seeks to classify human emotional states directly from the acoustic, prosodic, and spectral properties of voice signals. Historically, SER pipelines relied heavily on classical machine learning classifiers, such as Support Vector Machines (SVMs) and Hidden Markov Models (HMMs), trained on manually engineered acoustic feature sets [13,14]. However, the inherent non-stationarity of speech signals, compounded by environmental noise, has driven a paradigm shift toward advanced, data-driven signal decomposition paired with deep neural architectures. The enhanced robustness of modern deep learning SER models has facilitated their transition from experimental human-computer interaction to high-stakes clinical deployments.

Foundational clinical studies have demonstrated the efficacy of SER as a non-invasive diagnostic tool, such as in the detection of clinical depression [15]. Because humans cannot easily mask micro-variations in speech rate, acoustic energy, and spectral flatness, highly accurate SER systems are increasingly employed to detect mood disorders and monitor patient progress in smart environments [16]. This clinical necessity underscores the urgent demand for architectures that are not only highly accurate but computationally deterministic and robust against real-world noise.

Despite the push toward end-to-end models, extracting robust acoustic representations remains the vital prerequisite for high-accuracy SER. Standard spectral features, including MFCCs and basic spectrograms, provide a baseline representation of the vocal tract's shape and intonation [17]. Traditionally extracted using the Short-Time Fourier Transform (STFT), other fundamental signal processing techniques, such as the Constant-Q Transform (CQT) originally developed for music processing, have also been adapted to provide better logarithmic frequency resolution [18]. Regardless of the transform, raw features remain highly susceptible to mode mixing. Researchers previously explored Empirical Mode Decomposition (EMD); however, its recursive sifting process often resulted in severe boundary effects [19]. Consequently, the literature has heavily advocated for VMD, which has been successfully integrated into the SER pipeline as a non-recursive, adaptive filter bank [20]. By mathematically decomposing noisy signals into band-limited IMFs, VMD effectively isolates salient frequency bands [21]. Studies consistently demonstrate that extracting features directly from these optimized 1D sub-bands yields substantially higher classification accuracy compared to utilizing the raw global signal [7]. Recent literature has expanded upon VMD by integrating the Teager-Kaiser Energy Operator (TKEO) [22], permutation entropies [23], Multi-Resolution strategies (MRVMD) [24], and the Hilbert Transform [25]. Despite these advances in 1D processing, the structural separation of transient and continuous energy in the 2D time-frequency domain remains largely isolated from adaptive frequency methods. HP decomposition, achieved via orthogonal median filtering, separates continuous tonal structures from broadband transient noises [26].

As feature extraction has matured, the classification engines in SER have evolved into complex, hybrid spatial-temporal topologies. Because human emotion unfolds dynamically, isolated 2D CNNs often fail to capture sequential context, leading to the integration of Recurrent Neural Networks (RNNs) or Bidirectional Long Short-Term Memory (BiLSTM) units [27]. In these hybrid models, convolutional layers identify spectral peaks while recurrent layers model long-range dependencies [28]. A defining breakthrough in recent literature is the widespread adoption of Transformer architectures and self-attention mechanisms (such as Conformers), which process acoustic sequences in parallel and dynamically weigh specific audio frames [29,30]. Concurrently, hybrid quantum-classical models (Quantum Machine Learning) are being tested to achieve high classification accuracy while reducing trainable parameters [31].

A fundamental bottleneck in SER research is the limited scale of annotated emotional speech corpora. Deep learning models are highly prone to overfitting on these small datasets. To mitigate this, the field has aggressively adopted Self-Supervised Learning (SSL) [11]. Foundation models, including Wav2Vec 2.0 and HuBERT, are pre-trained on unlabelled audio to learn generalized acoustic embeddings [12]. Recent studies indicate these SSL representations drastically outperform traditional feature engineering and exhibit superior cross-corpus generalization [32]. However, SSL foundation models require immense computational resources, high inference latency, and severe memory overhead, limiting their viability for edge-device deployment.

Despite significant advancements in adaptive signal processing and deep neural architectures, current literature largely treats 1D time-domain decomposition (VMD) and 2D spatial separation (HP filtering) as mutually exclusive techniques. Furthermore, the increasing reliance on parameter-heavy Transformer models or massive SSL foundation embeddings restricts deployment in resource-constrained environments.

Motivated by these gaps, this research proposes a deterministic, lightweight ensemble framework. By concurrently extracting orthogonal affective cues across both 1D adaptive modes and 2D harmonic-percussive spatial maps, and fusing them via a dynamic attention mechanism, this study aims to partially mitigate complex high-arousal class confusion without the computational overhead of large-scale foundation models.

3 Methodology

The proposed DDE-SER represents a shift from heuristic feature stacking to mathematically structured, multi-domain signal decomposition. As illustrated in the schematic overview in Fig. 1, the architecture concurrently processes the acoustic input across the raw 1D signal level and the 2D transformed spectrogram level to extract orthogonal affective features. Following independent deep convolutional feature extraction within each branch, the resultant high-dimensional vectors are integrated via a trainable gated attention fusion mechanism to determine the final emotion class.

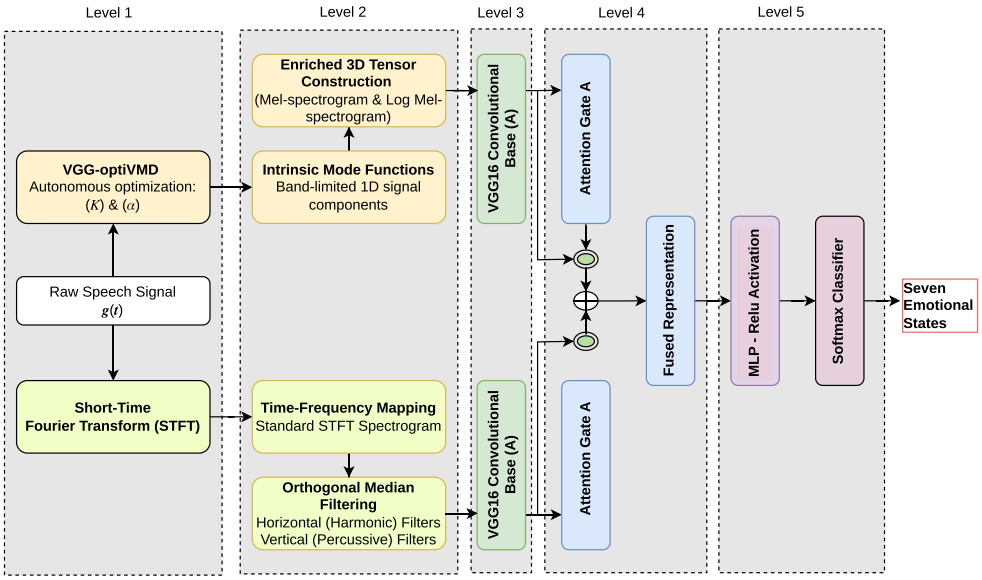


Figure 1: Schematic overview of the proposed DDE-SER framework, depicting the parallel extraction of 1D adaptive frequencies and 2D structural representations, followed by a gated attention fusion mechanism.

3.1 Branch A: Adaptive Mode Extraction via VGG-optiVMD

Human speech features dynamic frequency shifts that challenge standard filtering techniques. To address this, Branch A utilizes VMD as an adaptive filter bank to decompose the 1D audio waveform into band-limited IMFs. Because standard VMD requires manual tuning of the mode count (K) and penalty factor (α), we deploy the VGG-optiVMD algorithm. This approach leverages gradient feedback from a convolutional network to autonomously discover optimal hyperparameters, thereby preserving critical emotional cues without manual intervention. For an exhaustive mathematical formulation of VMD and the VGG-optiVMD optimization process, readers are referred to our prior foundational work [9].

3.2 Branch B: Structural Spectrogram Extraction (HP-Mel)

While Branch A targets 1D frequency bandwidths, Branch B is engineered to exploit 2D morphological differences. Tonal emotional cues (e.g., pitch inflections in sadness) manifest as continuous horizontal lines in a spectrogram, whereas broadband transient noises (e.g., explosive consonants in anger) appear as vertical ridges. The discrete signal $g(t)$ is initially transformed into a magnitude spectrogram $S(m,k)$ via the STFT.

To achieve structural separation, orthogonal median filters of length L_h (horizontal) and L_p (vertical) are applied to $S(m,k)$. The harmonic enhanced spectrogram $\tilde{S}_h$ and percussive enhanced spectrogram $\tilde{S}_p$ are subsequently isolated by computing Wiener-like soft masking matrices M_h and M_p:

$$M_h(m,k) = \frac{\tilde{S}_h^2(m,k)}{\tilde{S}_h^2(m,k) + \tilde{S}_p^2(m,k) + \varepsilon}, \quad M_p(m,k) = \frac{\tilde{S}_p^2(m,k)}{\tilde{S}_h^2(m,k) + \tilde{S}_p^2(m,k) + \varepsilon} \tag{1}$$

where ε prevents division by zero. The harmonic and percussive magnitudes are obtained via Hadamard multiplication: $H_{mag} = M_h \odot S$ and $P_{mag} = M_p \odot S$. The final 2D hybrid feature map F_{HP} is mapped to the Mel scale and logarithmically compressed:

$$F_{HP}(m,v) = \log_{10}\left(\Phi_{Mel}\left[\frac{H_{mag}(m,k) + P_{mag}(m,k)}{2}\right] + 1\right) \tag{2}$$

3.3 Deep Feature Extraction via Parallel CNNs

The resultant multi-dimensional tensors are fed into parallel VGG16 architectures acting as pure convolutional extractors (with the original classification heads removed). VGG16 was deliberately selected over parameter-heavy models like Transformers due to its shallow, sequential structure. Its localized receptive fields excel at capturing the specific edge-and-ridge morphologies generated by VMD and HP separation. This ensures the model's accuracy is driven by the mathematical orthogonality of the inputs rather than a black-box extractor. The final convolutional maps are flattened into semantic representation vectors $v_A, v_B \in \mathbb{R}^{2048}$.

3.4 Trainable Gated Attention Fusion Mechanism

Standard linear concatenation incorrectly assumes both decomposition branches are equally informative for every emotion. To grant representational flexibility, we implemented a Gated Attention Mechanism. Using trainable parameter matrices $W_{att}^A, W_{att}^B \in \mathbb{R}^{2048\times2048}$ and bias vectors b_{att}^A, b_{att}^B, the network computes an attention score vector for each branch via a Sigmoid activation function $\sigma(z) = (1 + e^{-z})^{-1}$:

$$e_A = \sigma(W_{att}^A v_A + b_{att}^A), \quad e_B = \sigma(W_{att}^B v_B + b_{att}^B) \tag{3}$$

These vectors act as informational filters. The final fused representation v_{fused} is generated through the Hadamard (element-wise) multiplication of the extracted features with their respective attention gates, followed by vector addition:

$$v_{fused} = (e_A \odot v_A) \oplus (e_B \odot v_B) \tag{4}$$

3.5 Algorithmic Workflow and Computational Complexity

During the forward pass, DDE-SER concurrently routes the raw audio through the 1D VMD and 2D HP pipelines. The extracted IMFs and spatial maps are processed by the parallel VGG16 networks, dynamically weighted by the Gated Attention mechanism, and classified via a multi-layer perceptron (MLP) using a Softmax function. Computationally, the framework's time complexity is firmly bounded by the Fast Fourier Transforms (FFTs) required in both the VMD, solved via the Alternating Direction Method of Multipliers (ADMM), and the STFT processes, scaling at $\mathcal{O}(N \log N)$ for an input signal of length N. Because the deep learning layers operate on fixed-size 128×128 representations, their processing overhead remains strictly $\mathcal{O}(1)$ relative to the audio length. This mathematical efficiency ensures the architecture avoids the exponential

parameter scaling of standard Transformer models, and the computational profile is consistent with edge deployment requirements for real-time affective computing.

4 Experimental Setup

4.1 Datasets

We utilized two highly benchmarked emotional speech corpora to evaluate the framework across different phonetic structures:

1. *Berlin Database of Emotional Speech (EMO-DB):* A German dataset comprising 535 acted utterances from 10 professional actors (5 male and 5 female). The specific breakdown consists of ten predefined everyday German sentences delivered across seven distinct emotional states (anger, boredom, disgust, fear, happiness, sadness, and neutral) [33].
2. *Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [34]:* A North American English dataset. We utilized the speech subset configured as a seven-class problem: the Calm class was merged into Neutral (reported as "Neutral (inc. Calm)"), and Surprised was retained as a separate class, yielding 2760 distinct audio files across seven emotion categories. The granular distribution features 24 professional actors (12 male, 12 female) delivering 2 specific statements at 2 varying levels of emotional intensity.

4.2 Data Preprocessing and Acoustic Augmentation

To counteract the overfitting risks inherent to small, acted datasets, we implemented a robust acoustic augmentation strategy using the `librosa` library [35]. Bypassing heuristic blind oversampling, we developed an in-place perturbation methodology simulating natural speech variability through three states: (1) Original Speech, (2) Pitch Shifting (upward by two half-steps), and (3) Dynamic Noise Injection (scaled to 0.5% of peak amplitude to mimic ambient interference). While acted corpora provide clean emotional baselines, the dynamic noise injection explicitly serves as a bridging mechanism to simulate unpredictable, real-world acoustic conditions, preparing the model for practical deployment. Furthermore, while drastic pitch manipulation can alter the perceived valence or arousal of an utterance, our pitch shifting was strictly limited to a micro-level variation of exactly two semitones. This specific boundary ensures the preservation of the emotional envelope, maintaining macro-level prosody, rhythm, and emotional intent, while effectively simulating the necessary inter-speaker anatomical variations in vocal tract length for generalized learning. This approach tripled the dataset sizes prior to extraction. To maintain absolute evaluation integrity, all augmented variants of a speaker's utterances were strictly confined alongside their original files in either the training or testing folds, ensuring zero data leakage.

4.3 Implementation Details

The framework was built using Python 3.10 and TensorFlow/Keras. The 1D VMD was executed via `vmdpy` (ADMM tolerance $1e^{-7}$), and the 2D HP filtering utilized `librosa` (STFT Hanning window 2048, hop length 512). Crucially, the VMD hyperparameters (K and α) are not heuristically initialized; instead, they are dynamically resolved per signal during training via the VGG-optiVMD feedback loop. The complete DDE-SER architecture, encompassing the dual-VGG16 convolutional bases, the Gated Attention mechanism, and the MLP classifier, comprises approximately 29.5 million trainable parameters. The network was optimized using the Adam optimizer (learning rate $1e^{-4}$) with early stopping set to 15 epochs. To ensure complete reproducibility, the full data curation pipeline, and the raw datasets are open-sourced and publicly available[1].

[1]Code and datasets are available at: https://github.com/DavidHason/dde-speech-emotion-recognition.

4.4 Speaker-Independent Cross-Validation (SICV) Strategy

A persistent flaw in SER literature is the use of random data splitting, which leaks speaker-specific traits into the validation set and artificially inflates accuracy. To rigorously test generalizability, we employed a strict SICV protocol. For a dataset containing N speakers, the model trains N separate times. In each fold, data from $N - 1$ speakers trains the model, leaving the single unseen speaker entirely isolated for testing. Because this protocol systematically evaluates every single speaker in the population as an isolated hold-out set, the variance of the model is inherently captured by the fold-to-fold performance differences. Therefore, rather than relying on randomized resampling techniques such as bootstrapping, we ensure statistical transparency by explicitly reporting the mean accuracy alongside the ±SD across all N folds. This approach provides a mathematically robust, deterministic measure of the framework's generalizability and stability.

5 Results and Discussion

5.1 Performance Analysis on EMO-DB

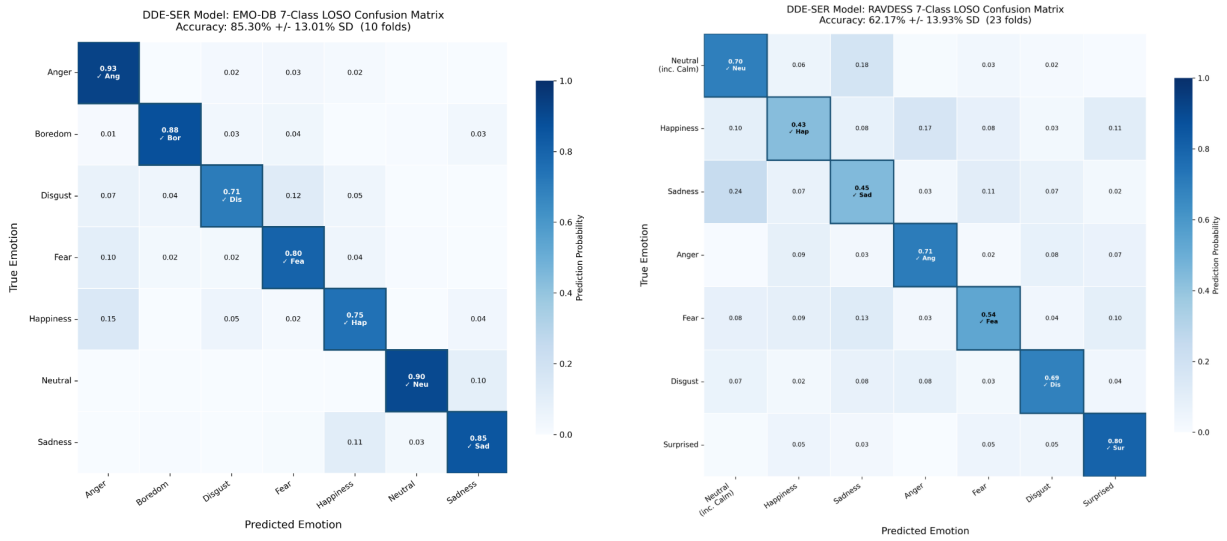
Under the rigorous SICV protocol, the framework achieved a mean overall accuracy of 85.30 ± 13.01% (SD) across all 10 held-out speaker folds on the augmented EMO-DB dataset, with fold-wise accuracy ranging from 58.50% to 99.05% (95% CI: [76.00%, 94.61%]; Table 1). Fold-wise overall accuracies (%) across the ten held-out speakers were: 58.50, 82.18, 91.47, 86.84, 92.73, 99.05, 96.17, 93.24, 85.71, and 67.14. The wide inter-fold range reflects genuine speaker-level variance; Fold 1 (58.50%) and Fold 10 (67.14%) represent speakers whose acoustic characteristics deviated most from the training population, whereas Fold 6 (99.05%) and Fold 7 (96.17%) yielded near-ceiling performance.

Table 1: Classification metrics and SICV statistical summary for the EMO-DB Dataset using SICV (7 classes, 10 speaker folds). AUC: One-vs.-Rest. SD reported with $df = 9$; 95% CI from t-distribution.

Emotion Class	Precision	Recall	F1-Score	AUC
Anger (Wut)	0.8445	0.9265	0.8836	0.9864
Boredom (Langeweile)	0.9425	0.8765	0.9083	0.9892
Disgust (Ekel)	0.7538	0.7101	0.7313	0.9757
Fear (Angst)	0.7905	0.8019	0.7962	0.9733
Happiness (Freude)	0.7644	0.7465	0.7553	0.9606
Neutral	0.9492	0.9032	0.9256	0.9962
Sadness (Trauer)	0.8559	0.8523	0.8541	0.9881
Macro Average	0.8430	0.8310	0.8364	0.9814
Weighted Average	0.8478	0.8467	0.8465	—
SICV Statistical Summary *(10 speaker folds, SICV)*				
Overall Accuracy—Mean ± SD			85.30 ± 13.01%	
95% Confidence Interval			[76.00%, 94.61%]	
Fold-wise Range (Min–Max)			58.50%–99.05%	
Macro AUC (One-vs.-Rest)			0.9814	

The per-class analysis confirms strong discriminability across the majority of emotions. Neutral achieved the highest F1-score (0.9256) and the highest Area Under the Curve (AUC) at 0.9962, demonstrating the VMD branch's capacity to isolate the narrow-band, steady-state energy characteristic of neutral speech.

Boredom also performed strongly (F1: 0.9083, AUC: 0.9892) and Anger (F1: 0.8836, AUC: 0.9864), confirming the HP branch's effectiveness at separating tonal continuity from broadband percussive energy. Disgust remained the most challenging class (F1: 0.7313, AUC: 0.9757), consistent with its broad intra-class acoustic variance. Happiness exhibited moderate performance (F1: 0.7553, AUC: 0.9606), with residual confusion attributable to its spectral proximity to Anger in the high-arousal quadrant of Russell's Circumplex Model. Notably, the macro AUC of 0.9814 across all seven classes confirms strong probabilistic class separability even for the more challenging emotions evident in Fig. 2a. While the DDE-SER ensemble partially mitigates misclassifications across the broader emotional spectrum, the persistent confusion between acoustically proximate high-arousal states underscores that full valence separation remains an open challenge for acoustic-only models.



(a) EMO-DB Dataset

(b) RAVDESS Dataset

Figure 2: Normalized SICV confusion matrices. (**a**) EMO-DB dataset: Anger (0.93) and Neutral (0.90) achieve the highest recall; Disgust (0.71) and Happiness (0.75) exhibit residual misclassifications into adjacent high-arousal states. (**b**) RAVDESS 7-class dataset, illustrating the dispersive misclassification patterns for Happiness and Sadness alongside the high recall achieved for Surprised (0.80) and Anger (0.71).

5.2 Performance Analysis on RAVDESS

The English-language RAVDESS evaluation was conducted under the seven-class SICV protocol (Calm merged into Neutral; Surprised retained), with 23 held-out speaker folds. The framework achieved a mean overall accuracy of 62.17 ± 13.93% (SD) across all 23 folds, with fold-wise accuracy ranging from 35.83% to 92.50% (95% CI: [56.15%, 68.20%]; Table 2). AUC values in Table 2 are One-vs.-Rest, aggregated across all 23 SICV folds. Fold-wise overall accuracies (%) were: 60.00, 92.50, 72.50, 60.00, 70.00, 65.00, 84.17, 79.17, 54.17, 64.17, 48.33, 61.67, 35.83, 47.50, 47.50, 81.67, 60.83, 60.00, 59.17, 53.33, 70.83, 40.83, and 60.83.

Table 2: Classification metrics and SICV statistical summary for the RAVDESS dataset using SICV (7 classes: Calm merged into Neutral, Surprised retained; 23 speaker folds). AUC: One-vs.-Rest. SD reported with $df = 22$; 95% CI from t-distribution.

Emotion Class	Precision	Recall	F1-Score	AUC
Neutral (inc. Calm)	0.6754	0.6975	0.6863	0.9362
Happiness	0.5096	0.4321	0.4676	0.8349
Sadness	0.4220	0.4484	0.4348	0.8443
Anger	0.6832	0.7092	0.6960	0.9379
Fear	0.6099	0.5353	0.5702	0.8954
Disgust	0.7056	0.6902	0.6978	0.9341
Surprised	0.6991	0.8016	0.7468	0.9572
Macro Average	0.6150	0.6163	0.6142	0.9057
Weighted Average	0.6190	0.6217	0.6190	—
SICV Statistical Summary *(23 speaker folds, SICV)*				
Overall Accuracy—Mean ± SD			62.17 ± 13.93%	
95% Confidence Interval			[56.15%, 68.20%]	
Fold-wise Range (Min–Max)			35.83%–92.50%	
Macro AUC (One-vs.-Rest)			0.9057	

The per-class analysis reveals a clear gradient of acoustic discriminability. Surprised achieved the highest recall (0.8016), attributable to its distinctive explosive onset transients isolated by the HP branch. Anger also performed strongly (recall 0.7092), consistent with the EMO-DB findings. The most challenging classes were Happiness (recall 0.4321) and Sadness (recall 0.4484). Crucially, their low recall does not arise from mutual confusion; Happiness and Sadness occupy diametrically opposed quadrants of Russell's Circumplex Model, but from two distinct intra-class failure modes evident in Fig. 2b. Sadness is misclassified predominantly as Neutral (inc. Calm) (0.24 of instances), reflecting the inherently low-energy, low-arousal acoustic profile it shares with the merged Neutral/Calm category, whose boundary is consequently the most porous in the matrix. Happiness, by contrast, disperses across the high-arousal cluster, with its principal leakage directed toward Anger (0.17) and Surprised (0.11), states whose elevated pitch and energy envelopes overlap with its own. Despite these confusions, the macro One-vs.-Rest AUC of 0.9057 confirms strong class separability in the probabilistic space. Consequently, while the dual-domain approach partially mitigates feature crossover for acoustically structured emotions such as Anger and Surprised, the residual errors for Happiness and Sadness stem from two separable sources—high-arousal spectral overlap and low-arousal energy ambiguity—each of which remains an open challenge in acoustic-only SER.

5.3 Comparison with State-of-the-Art Approaches

We benchmarked the framework against recent state-of-the-art (SOTA) SER methodologies under isolated SICV validation protocols (Table 3). While some recent literature reports accuracy metrics exceeding 85%–90% within their abstracts [22–24], those headline figures rely on random data splitting techniques highly prone to speaker leakage; only SICV figures are used here for equitable comparison. On EMO-DB, DDE-SER achieves 85.30%, representing an 11.50% absolute margin over the closest SICV baseline, confirming the benefit of the dual-decomposition and gated fusion approach under strict zero-leakage evaluation. For RAVDESS, the proposed framework was evaluated under a seven-class protocol (Calm merged

into Neutral; Surprised retained), whereas the referenced baselines used different emotion subsets; the macro AUC of 0.9057 confirms strong discriminative capacity within this configuration. This performance differential supports our core hypothesis: while adaptive 1D signal decompositions are effective at capturing non-stationary energy shifts, fusion with 2D structural representations via dynamic attention gating further improves the resolution of fine-grained acoustic overlaps.

Table 3: Comparison of DDE-SER accuracy against recent state-of-the-art models utilizing isolated SICV validation.

Study	Methodology	EMO-DB Acc.	RAVDESS Acc.
Ravi and Taran [22]	VMD + TKEO + SVM	71.30%	69.80%
Mishra et al. [23]	VMD + ApEn/PrEn + DNN	72.10%	71.40%
Mishra et al. [24]	MRVMD + Deep Features	73.50%	74.20%
Mishra et al. [25]	VMD + Hilbert Transform	73.80%	—
Proposed	**DDE-SER (VMD + HP + Attention)**	**85.30%**	**62.17%***

Note: Bold entries identify the proposed framework; they do not denote the highest value in a column. *RAVDESS result for DDE-SER uses a seven-class configuration (Calm → Neutral; Surprised retained); baselines used different class subsets.

5.4 Limitations

Several important scope boundaries must be acknowledged. First, all experiments were conducted exclusively on acted, studio-quality corpora (EMO-DB and RAVDESS). Validation on spontaneous, naturalistic speech, where emotional expression is conversational, partially masked, or acoustically degraded, has not been performed. Consequently, generalisation to spontaneous corpora such as IEMOCAP cannot be assumed from the present results. Second, while the dynamic noise injection augmentation is designed to partially simulate real-world acoustic interference, the model has not been tested under genuine in-the-wild or noisy field conditions. Third, no cross-corpus evaluation has been conducted; performance on speaker populations, languages, or recording conditions outside those covered by EMO-DB and RAVDESS remains unknown. Fourth, the RAVDESS evaluation adopts a seven-class configuration (Calm merged into Neutral; Surprised retained) that differs from the class subsets used in several prior studies, limiting direct numerical comparability in the SOTA table. These boundaries do not diminish the validity of the controlled evaluation presented here, but they do define the limits within which the reported accuracies should be interpreted. Addressing these gaps through spontaneous and cross-corpus evaluation is the primary directive of future work.

6 Conclusion and Future Work

This research introduces the DDE-SER, an architecture integrating 1D VMD with 2D HP spectrogram representations to extract orthogonal affective cues. Building upon prior individual extraction methodologies [9,10], the primary novel contribution is the Gated Attention fusion mechanism that bridges these established pipelines. By dynamically weighting the two orthogonal feature branches, the model effectively diminishes representational redundancy and partially mitigates high-arousal emotional class confusion, achieving 85.30 ± 13.01% (SD; macro AUC: 0.9814) on EMO-DB and 62.17 ± 13.93% (SD; macro AUC: 0.9057) on the seven-class RAVDESS corpus under rigorous SICV testing. As noted in the Limitations section, evaluation has been restricted to acted, controlled-recording datasets with a specific RAVDESS class configuration, and validation on spontaneous, noisy, or cross-corpus speech remains untested.

Future exploration will prioritize evaluating DDE-SER on spontaneous, naturalistic datasets (e.g., IEMOCAP) to assess its viability beyond controlled environments. Additionally, to improve classification

accuracy on highly variant classes like Happiness and Sadness, subsequent iterations will explore advanced domain-specific data augmentations, such as synthetic minority oversampling via Generative Adversarial Networks (GANs) or dynamic SpecAugment. Aligning the RAVDESS evaluation with the class configurations used in the comparative literature will also be pursued to enable fully equitable SOTA benchmarking. Finally, adopting robust data documentation frameworks will standardize acoustic metadata tracking, facilitating cross-corpus generalization in affective computing.

Acknowledgement: The authors would like to acknowledge the use of the EMO-DB and RAVDESS databases for providing the speech data necessary for these experiments. The authors acknowledge the use of generative AI tools to improve the readability and grammatical flow of this manuscript. The authors have reviewed and edited the content and take full responsibility for the final text.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization and methodology, David Hason Rudd; writing, original draft preparation, David Hason Rudd; writing, review and editing, Huan Huo, Md Rafiqul Islam, Xianzhi Wang and Cesar Sanin; supervision, Huan Huo. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available. EMO-DB and RAVDESS are publicly accessible databases.

Ethics Approval: Not applicable. This study utilized publicly available, anonymized datasets (EMO-DB and RAVDESS) and did not involve direct human or animal subjects.

Conflicts of Interest: Given his role as guest editor of this journal, Md Rafiqul Islam had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

References

1. Mustaqeem, Kwon S. A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors. 2020;20(1):183. doi:10.3390/s20010183.
2. Koduru A, Valiveti HB, Budati AK. Feature extraction algorithms to improve the speech emotion recognition rate. Int J Speech Technol. 2020;23:45–55.
3. Basharirad B, Moradhaseli M. Speech emotion recognition methods: a literature review. AIP Conf Proc. 2017;1891(1):020105.
4. Gumelar AB, Purnomo MH, Widodo A, Kurniawan A, Yuniarno EM, Kristanto AA, et al. Human voice emotion identification using prosodic and spectral feature extraction based on deep neural networks. In: 2019 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM). Piscataway, NJ, USA: IEEE; 2019. p. 1–8.
5. Popova AS, Rassadin AG, Ponomarenko AA. Emotion recognition in sound. In: International Conference on Analysis of Images, Social Networks and Texts. Cham, Switzerland: Springer; 2017. p. 225–31.
6. Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms. In: Proceedings of Interspeech 2017; 2017 Aug 20–24; Stockholm, Sweden. p. 1089–93.
7. Dendukuri LS, Hussain SJ. Emotional speech analysis and classification using variational mode decomposition. Int J Speech Technol. 2022;25(2):1–13. doi:10.1007/s10772-022-09970-z.
8. Deb S, Dandapat S, Krajewski J. Analysis and classification of cold speech using variational mode decomposition. IEEE Trans Affect Comput. 2020;11(2):296–307. doi:10.1109/taffc.2017.2761750.

9. Rudd DH, Huo H, Xu G. An extended variational mode decomposition algorithm developed speech emotion recognition performance. In: Advances in knowledge discovery and data mining (PAKDD 2023). Cham, Switzerland: Springer; 2023. p. 219–31.
10. Rudd DH, Huo H, Xu G. Leveraged Mel-spectrograms using harmonic and percussive components in speech emotion recognition. In: Advances in knowledge discovery and data mining (PAKDD 2022). Cham, Switzerland: Springer; 2022. p. 392–404.
11. Neumann M, Vu NT. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019 May 12–17; Brighton, UK.
12. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst. 2020;33:12449–60.
13. Weninger F, Wöllmer M, Schuller B. Emotion recognition in naturalistic speech and language, a survey. In: Emotion recognition: a pattern analysis approach. New York, NY, USA: John Wiley & Sons; 2015. p. 237–67.
14. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. 2011;44(3):572–87.
15. Low LS, Maddage NC, Lech M, Sheeber L, Allen N. Detection of clinical depression in adolescents' speech during family interactions. IEEE Trans Biomed Eng. 2011;58(3):574–86. doi:10.1109/tbme.2010.2091640.
16. Xu X, Fu C, Camacho D, Park JH, Chen J. Internet of Things for emotion care: advances, applications, and challenges. Cogn Comput. 2024;16(6):2812–32. doi:10.1007/s12559-024-10327-8.
17. Issa D, Demirci MF, Yazici A. Speech emotion recognition with deep convolutional neural networks. Biomed Signal Process Control. 2020;59:101894. doi:10.1016/j.bspc.2020.101894.
18. Schörkhuber C, Klapuri A. Constant-Q transform toolbox for music processing. In: 7th Sound and Music Computing Conference; 2010 Jul 21–24; Barcelona, Spain.
19. Wu Z, Huang NE. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Anal. 2009;1(1):1–41. doi:10.1142/s1793536909000047.
20. Dragomiretskiy K, Zosso D. Variational mode decomposition. IEEE Trans Signal Process. 2014;62(3):531–44.
21. Lal GJ, Gopalakrishnan E, Govind D. Epoch estimation from emotional speech signals using variational mode decomposition. Circuits Syst Signal Process. 2018;37(8):3245–74. doi:10.1007/s00034-018-0804-x.
22. Ravi, Taran S. A nonlinear feature extraction approach for speech emotion recognition using VMD and TKEO. Appl Acoust. 2023;214(2):109667. doi:10.1016/j.apacoust.2023.109667.
23. Mishra SP, Warule P, Deb S. Variational mode decomposition based acoustic and entropy features for speech emotion recognition. Appl Acoust. 2023;212(5):109578. doi:10.1016/j.apacoust.2023.109578.
24. Mishra SP, Warule P, Deb S. Improvement of emotion classification performance using multi-resolution variational mode decomposition method. Bio Sig Proc Cont. 2024;89:105708.
25. Mishra SP, Warule P, Deb S. Speech emotion recognition using a combination of variational mode decomposition and Hilbert transform. Appl Acoust. 2024;222(5):110046. doi:10.1016/j.apacoust.2024.110046.
26. Fitzgerald D. Harmonic/percussive separation using median filtering. In: 13th International Conference on Digital Audio Effects (DAFx); 2010 Sep 6–10; Graz, Austria.
27. Zhao J, Mao X, Chen L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomed Signal Process Control. 2019;47(4):312–23. doi:10.1016/j.bspc.2018.08.035.
28. Chernykh V, Prikhodko P. Emotion recognition from speech with recurrent neural networks. arXiv:1701.08071. 2017.
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA.
30. Gulati A, Qin J, Chiu CC, Parmar N, Zhang Y, Yu J, et al. Conformer: convolution-augmented transformer for speech recognition. arXiv:2005.08100. 2020.
31. Yang C, Chen J, Li X. Quantum machine learning for speech emotion recognition: a preliminary study. IEEE Trans Quantum Eng. 2025;6:1–12.

32. Fu H, Li Q, Tao H, Zhu C, Xie Y, Guo R. Cross-corpus speech emotion recognition based on causal emotion information representation. IEICE Trans Inf Syst. 2024;107(8):1097–100. doi:10.1587/transinf.2023edl8087.
33. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. In: Proceedings of Interspeech 2005; 2005 Sep 4–8; Lisbon, Portugal. p. 1517–20.
34. Livingstone SR, Russo FA. The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One. 2018;13(5):e0196391.
35. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, et al. librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference; 2015 Jul 6–12; Austin, TX, USA. p. 18–25.