Speech Emotion Recognition using Mel Spectrogram HPCA and Variational Mode Decomposition

1st David Hason Rudd
Faculty of Engineering and IT
The University of Technology Sydney
Sydney, Australia
david.hasonrudd@uts.edu.au

2nd Xingyi Gao
Faculty of Engineering and IT
The University of Technology Sydney
Sydney, Australia
Xingyi.Gao@student.uts.edu.au

3rd Md Rafiqul Islam
Information Systems (Data Analytics)
Australian Institute of Higher Education (AIH)
Sydney, Australia
r.islam@aih.edu.au

2nd Huan Huo
Faculty of Engineering and IT
The University of Technology Sydney
Sydney, Australia
huan.huo@uts.edu.au

4rd Guandong Xu
Faculty of Engineering and IT
Advanced Analytics institute (AAi)
Sydney, Australia
guandong.xu@uts.edu.au

***Abstract*—The rapid evolution of affective computing demands sophisticated methodologies to enhance the reliability and effectiveness of speech emotion recognition (SER). This study integrates harmonic-percussive component analysis (HPCA) with variational mode decomposition (VMD) to overcome various drawbacks for conventional speech emotion recognition (SER) methodologies that primarily rely on stand-alone feature extraction techniques. This implementation refines acoustic feature extraction and optimizes VMD decomposition to prevent information loss from mode duplication and mixing problems. We propose a feature map generator that channels the enhanced feature vectors into a convolutional neural network, specifically the VGG16 model, and the model is further enriched by incorporating diverse acoustic features including HP and log Mel spectrograms into two-dimensional spaces to intensify data augmentation and enrich emotional feature representation. Extensive testing on Berlin EMO-DB and RAVDESS databases confirmed positive impacts for the proposed HP-VMD model performance, achieving robust classification accuracy of 96.67%. Thus, the proposed integrated approach to developing SER systems significantly enhances empathetic human computer interactions.**

***Index Terms*—Speech emotion recognition, Harmonic-percussive component analysis, Variational mode decomposition, Mel spectrogram, Convolutional neural networks, Acoustic features, Human computer interaction**

I. Introduction

Speech emotion recognition (SER) is pivotal to refine human computer interaction (HCI) across diverse domains including customer service, health care, security, and entertainment. SER systems identify human emotions from voice characteristics that indicate the speaker's emotional state, enhancing service delivery effectiveness for applications ranging from call centers to therapeutic diagnostics [1].

Traditional SER methodologies primarily utilize prosodic, acoustic, and linguistic features extracted from speech signals to decode emotions. Among these features, Mel frequency cepstral coefficients (MFCCs), chromagrams, and Mel spectrograms have shown significant efficacy in capturing emotional content embedded in speech. However, these systems often face challenges related to the nonstationary nature of speech signals, where conventional methods such as short time Fourier transform and empirical mode decomposition may not adequately capture subtle dynamics for emotional expressions due to limitations handling overlapping frequencies and mode mixing [2].

To address these challenges, this study proposes integrating HPCA with VMD to enhance feature extraction and classification processes for SER. HPCA effectively isolates the harmonic and percussive elements of speech, providing a clear delineation between voice tone and temporal dynamics, which are crucial for emotion recognition; whereas VMD decomposes nonstationary signals into adaptively determined modes with minimal overlap. This, combining HPCA and VMD provides a more precise extraction of emotional features.

The proposed method leverages HPCA and VMD strengths by constructing a robust feature map to feed into the VGG16 convolutional neural network (CNN) model, adjusted for SER. This integrated approach achieves superior classification accuracy and model generalizability across different emotional states and datasets by optimizing decomposition parameters and enhancing signal representation.

Significant contributions from this study can be summarized as follows.

- Integrating HPCA and VMD to enhancing precision in emotional state classification from speech.
- As far as we are aware, this study offers a novel approach to utilize HPCA and VMD as a dynamic method for acoustic feature augmentation, significantly enhancing SER system performance.

- Preliminary results indicate the proposed approach significantly outperforms current models in terms of accuracy and reliability on standard SER benchmarks, such as the Berlin EMO-DB database, achieving state-of-the-art results.

The remainder of this paper is organized as follows: Section II reviews recent advancements and related studies in SER. Section III elaborates on the methodologies employed using the proposed HP-VMD algorithm. Section IV presents experimental analyses, modeling, and results. Section V concludes the paper and outlines future research directions in real-world applications.

II. Recent Works

Previous SER studies have leveraged various methodologies to enhance emotion detection robustness and accuracy from speech signals. SER dynamic nature requires integrating diverse signal processing techniques to address intrinsic challenges associated with the non-stationary and complex nature of human speech. Human speech complexity is profoundly shaped by cultural values, traditions, and geographical isolation, leading to dialectical variations and linguistic borrowing influenced by social interactions and environmental adaptations. This linguistic diversity reflects the intricate interplay between cultural practices and geographical settings [20]. This section discusses foundational work in HPCA and VMD domains and their integration in SER.

Traditional SER approaches predominantly focused on extracting meaningful features from speech using methods, such as MFCCs, chromagrams, and spectrograms. Previous studies have explored various classification models, ranging from support vector machines (SVMs) and neural networks to more complex structures integrating convolutional layers [5]. For example, Huang et al. proposed a hybrid model combining deep CNNs with SVM, achieving significant emotion classification improvement from the EMO-DB database [6].

Similarly, HPCA has been explored to effectively separate harmonic and percussive elements from audio signals, providing clear distinctions beneficial for understanding tonal and rhythmic speech components [7]. However, the technique has not been widely integrated with other decomposition techniques in SER, which could potentially enhance discriminatory power for extracted features.

Variational mode decomposition decomposes nonstationary signals into their constituent modes, and has become increasingly applied across fields requiring detailed signal analysis beyond traditional Fourier methods. VMD can handle overlapping frequencies and adaptably select the number, making it particularly suitable for complex signal environments such as human speech [9]. In particular, VMD has been employed with SER to refine feature extraction processes, particularly when combined with machine learning classifiers to improve emotional accuracy [10], [14], [31].

Pandey and Seeja advance EEG-based emotion recognition by integrating VMD with deep neural networks. Their approach, which utilizes intrinsic mode functions for feature extraction, excels in subject-independent emotion classification, enhancing the system's applicability in diverse real-world scenarios without prior individual data [25]. Taran introduced a novel SER methodology combining VMD with the teager-kaiser energy operator. This integration enhances speech signal decomposition and feature extraction, significantly boosting classification performance with SVMs variants across varied emotional states [28].

Despite ongoing separate HPCA and VMD advances they have rarely been combined for SER. However, integrating these methodologies could potentially overcome various limitations with traditional approaches, including losing important temporal or spectral information and inability to effectively separate overlapping emotional cues in speech. Therefore, the current study proposes to address this gap by developing a hybrid feature extraction framework that leverages HPCA and VMD strengths. This integration will enhance extraction and classification of emotional states from speech signals, providing a more robust and accurate SER system that could significantly advance current practices.

Convergence these methodologies in a unified SER framework represents a novel approach in the field. Leveraging precise frequency decomposition from VMD with detailed harmonic and percussive separation from HPCA will help set a new benchmark in emotion recognition accuracy and reliability.

III. Methodology

This section outlines the methodologies utilized to integrate Mel spectrogram HPCA and VMD to enhance speech emotion recognition. The proposed integrated approach leverages both methodologies strengths to improve robustness and accuracy for emotion detection from speech signals.

A. Mel spectrogram feature extraction

The Mel spectrogram is utilized due to its effectiveness in representing the spectral energy distribution of speech signals over time, enabling a more nuanced analysis of the tonal and rhythmic elements critical for emotion recognition within the framework of HPCA. A major challenge for speech feature analysis is extracting pivotal characteristics from voice samples to form comprehensive vectors, ensuring vector size uniformity while preserving critical data. We consider speech acoustic dynamics by harnessing the analytical power of Mel spectrograms, which are intrinsic to various applications, including sound event detection, speaker and speech recognition, leveraged through a hybrid feature engineering process to maximize the classifier predictive accuracy [11], [12].

This study adopting the Mel scale, relating perceived to actual frequency, to accommodate human non-linear auditory perception. This significantly aids constructing an auditory-sensible representation for sound. The Mel frequency f_{mel} can be expressed as

$$f_{mel} = 2595 \cdot \log(1 + \frac{f}{700\text{Hz}}). \tag{1}$$

The proposed methodology designates 128 filter banks for feature extraction to achieve optimal frame size alignment. Sections III-F and IV show that voice signals are first digitized at 88 kHz sample rate using a Hanning window function [13], and the Mel spectrogram is subsequently constructed through strategic windowing and applying Mel filter banks, resulting in a coherent series of fast Fourier transforms.

The first dimension for the proposed feature map generation is the Mel spectrogram,

$$S(n,k) := \sum_{r=0}^{N-1} s(r+nH) \cdot \omega(r) \cdot e^{-\frac{j2\pi kn}{N}}, \tag{2}$$

where S is the spectrogram for signal s; $\omega : [0 : N-1]$ is a sine window function indicative of the window's span, H is stride length, n is the sequence number for the current frame, and N is the total point count for the discrete Fourier transformation.

This Mel spectrogram is subsequently fused with its calculated second dimension, yielding a two-dimensional (2D) feature map with size (128×128×2), which distinctly differentiates emotional amplitudes and frequencies.

A hybrid feature map is then calculated by decomposing the Mel spectrogram into harmonic $\widehat{H}$ and percussive $\widehat{P}$ components using a horizontal and vertical median filtering technique,

$$\widehat{H} = \widehat{S} \bigotimes M_H \quad \text{and} \quad \widehat{P} = \widehat{S} \bigotimes M_P. \tag{3}$$

Finally, the HPCA feature vector is obtained for each component,

$$\mathcal{HPC} = \frac{(\widehat{H} + \widehat{P})}{2}, \tag{4}$$

Provided enriched input for the VMD algorithm by separating tonal and dynamic elements from speech, hence enhancing VMD's capability to isolate emotion related features, and subsequently improving SER accuracy.

B. Harmonic-percussive component analysis

The proposed approach employs HPCA to dissect the speech signal into harmonic (tonal) and percussive (rhythmic) components. This separation is crucial to allow nuanced analysis for speech tonal quality and rhythmic intensity, which indicate emotional states. Emotions can influence pitch and timbre, manifesting distinctly in harmonic patterns; as well as rhythmic expressions, detectable in percussive elements. The process applies a median filtering technique to the Mel spectrogram for the speech signal, emphasizing distinct qualities for the harmonic and percussive structures. This proposed approach enhances speech feature clarity which is critical for effective emotion recognition [7].

C. Variational mode decomposition feature augmentation

The proposed approach employs VMD to decompose the speech signal into a predefined number of band-limited intrinsic mode functions (IMFs). This technique is particularly effective for handling non-stationary signals, such as human speech. The VMD process adopts an iterative approach to minimize the decomposed modes bandwidths, constrained by reconstructing the original signal from these modes. Optimization is typically achieved using the alternate direction method of multipliers (ADMM) to ensure accurate frequency based feature extraction essential for detecting emotional nuances in speech [8].

D. Integrating HPCA and VMD to enhance SER

The proposed approach integrates HPCA and VMD, leveraging their respective strengths, to significantly enhance speech emotion recognition. This is achieved through a feature fusion technique that first decomposes the Mel spectrogram harmonic and percussive components using HPCA, then reshapes the decomposed component data frames into concatenated feature vectors as input data for the VMD algorithm.

This integration enhances modal clarity and enriches informative features from the speech signal that indicate emotional states. Processing the concatenated and average vector value of harmonic and percussive components through VMD feature augmentation and VGG16 framework effectively increases emotional feature resolution and distinctiveness.

The HPCA output provides enriched inputs for the VMD algorithm, hence obtaining high accuracy emotion classification. This integration allows VMD and VGG networks to leverage temporal information provided by the HPCA, leading to more robust and comprehensive acoustic feature extraction. The proposed method provides a new approach to ensure precision and reliability of emotion recognition from speech, improving prediction accuracy and ensuring the system is less susceptible to common issues, e.g. overfitting, enhancing its applicability in real-world scenarios where emotional recognition is critical.

E. Proposed HP-VMD algorithm

The essence of our proposed method is the HP-VMD algorithm, which enhances extracting and classifying emotional states from speech by integrating HPCA with VMD. Starting with a preprocessed voice signal from the EMO-DB and RAVDESS databases, the algorithm is calibrated with at 88200 Hz sample rate and 2048 HOP window with 128×128 band and frame, providing a high-resolution basis for extracting acoustic features. Key parameters number of modes K, bandwidth control parameter α, and convergence tolerance `tol` (finetuned to $1e-9$), were initialized to facilitate accurate feature decomposition.

The algorithm commences with HPCA to extract harmonic and percussive components from the Mel spectrogram, capturing transient and sustained characteristics from the speech signal carrying emotional and tone data, respectively. These HP components are then incorporated into the VMD process, where the signal is decomposed into a predefined number of IMFs. Modes $\hat{g}_k$ and corresponding center frequencies $\hat{\omega}_k$ are iteratively refined using an Weiner filtering within the ADMM framework [24], ensuring that each IMF or sub-signal distinctly embodies specific emotional characteristics.

The HP-VMD algorithm optimizes parameter values for K and α to maximize SER classification accuracy and F1 score. The optimization is driven by a feedback loop that records performance metrics, tuning the hyperparameters until the algorithm converges on a feature set offering the most representative emotional content for accurate classification. Therefore, the proposed HP-VMD algorithm dissects a voice signal into fundamental emotional frequencies by fusing HPCA and VMD.

Algorithm 1: Proposed HP-VMD algorithm

Input: $g(t)$ preprocessed voice data.

Output: Emotion class id.

Initialization: HOP window size = 2048 with (128 Ã 128) band and frame; SR= 88200; modes K and α; tol=1e-9, DC=0, init=1, and τ=0; Convergence criterion τ tolerance; $\left\{\hat{g}_k^1\right\}, \left\{\hat{\omega}_k^1\right\}, \hat{\lambda}^1$; $n = 0$.

HPCA: Extract harmonic $\widehat{H}$ and percussive $\widehat{P}$ features from the Mel spectrogram $S(n,\ k)$;

1: $S(n,\ k) \ := \ \sum_{r=0}^{N-1} s(r + nH) \ . \ \omega(r) \ . \ e^{(\frac{-j2\pi.k.n}{N})}$

2: Median filter in horizontal and vertical directions:
$\widehat{H} = \widehat{S} \ \bigotimes \ M_H$
$\widehat{P} = \widehat{S} \ \bigotimes \ M_P$

3: $HPC = \hat{g}_k$ obtained by $\hat{g}_k = \frac{(\widehat{H}+\widehat{P})}{2}$

Repeat:

4: $n = n + 1$,

5: **for** k=1 : K **do**

6: update $\hat{g}_k$ for all $\omega \geq 0$:

$$\hat{g}_k^{n+1}(\omega) = \frac{\hat{g}(\omega) - \sum_{i<k} \hat{g}_i^{n+1}(\omega) - \sum_{i>k} \hat{g}_i^n(\omega) + \frac{\hat{\lambda}^n(\omega)}{2}}{1 + 2\alpha\left(\omega - \omega_k^n\right)^2}$$

update ω_k by using:

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega \left|\hat{G}_k(\omega)\right|^2 \, \mathrm{d}\omega}{\int_0^\infty \left|\hat{G}_k(\omega)\right|^2 \, \mathrm{d}\omega}$$

7: **end for**

8: Upgrade Lagrangian multiplier λ for dual accent $\forall\omega \geqslant 0$:

$$\lambda^n(\omega) = \lambda^n + \tau(g(\omega) - \sum_k g_k^{n+1}(\omega))$$

Until:

9: convergence: $\sum_{k=1}^K \|\hat{g}_k^{n+1} - \hat{g}_k^n\|_2^2 / \|\hat{g}_k^n\|_2^2 < \in$.

10: **return** Decomposed $g(t)$:$\{g_1(t), g_2(t), \ldots, g_K(t)\}$= IMFs; subtract all sub-signals

11: Set Parameters τ=0; DC=0; init=1; tol=1e-9; K=2; α=2000.

12: Record training set accuracy and F1 score in VGG16 classifier in optimum set of K and α

13: **while** max(ACC) **do**

do**if** ACC==max; $\alpha \leq 6000$; $K \leq 8$ **then**

15: Obtain optimum value of K and α.

16: **else**

$K = K + 1$; α=α+1000 go to step 3

17: **end if**

18: **end while**

F. Modelling

The proposed modeling strategy prioritizes both enriching feature vectors and mitigating overfitting by employing data augmentation, where feature vectors obtained from input signal $g(t)$ (defined in Algorithm 1) are decomposed into multiple modes. This augmentation expands the dataset and enhances model generalization capabilities. The HP-VMD algorithm is central to the proposed approach, extracting emotionally relevant information from speech signals. Optimal number of modes K and decomposition parameter α are determined iteratively, with experimental results guiding selecting $K \in [3, 8]$ and $\alpha \in [1000, 6000]$ for superior classification accuracy.

The hybrid model architecture leverages a modified CNN-VGG16 network for dynamic feature extraction. VGG16's inherent ability to identify subtle patterns, originally developed for image analysis, is effectively adapted for speech emotion recognition, and the extracted features are then classified by a flattening layer. The VGG fully connected layer is carefully built to prevent overfitting and promote non-linear mapping, combining activation functions ReLU, SELU, and TanH, with dropout regularization.

The model is trained using the ADAM optimizer with learning rate = 0.0001. The network architecture includes six fully connected hidden layers and training proceeds for 50 epochs with batch size = 4. A Softmax output function is used for classification. Hyperparameters for VGG16 are selected to balance computational efficiency with classification performance.

Figure 1 shows the proposed HP-VMD model architecture for enhanced SER. The proposed framework begins with data preprocessing, loading voice files with 88200 Hz sample rate, then segmented using a 2048 HOP window, which is the number of samples between successive frames. HP components are then extracted using the HPCA block and fed into the feature map for scaling. The core HP-VMD algorithm, VMD feature augmentation, takes these enriched inputs to further refine feature extraction, employing VMD as a dynamic method for acoustic feature augmentation. The CNN-VGG16 network is then utilized with multiple dense layers to fine-tune emotional data processing before classifying into seven emotion classes: anger, boredom, happiness, neutral, disgust, sadness, and fear.

IV. Experiment analysis

Experiments utilized the Librosa (Python) HPSS toolkit to extract acoustic features from the Berlin EMODB database [29] and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [30]. Both are comprehensive emotion databases, hence we used only these two databases since the intrinsic nature of the human voice is consistent. Therefore, experiments focused on examining the proposed hybrid feature extraction framework's performance.

Preprocessing quantized voice samples from the databases with 2048 HOP window size, length = 256, and 88200 Hz sample frequency to sharpen frequency details and minimize spectral leakage. Section IV-A shows the outcomes from

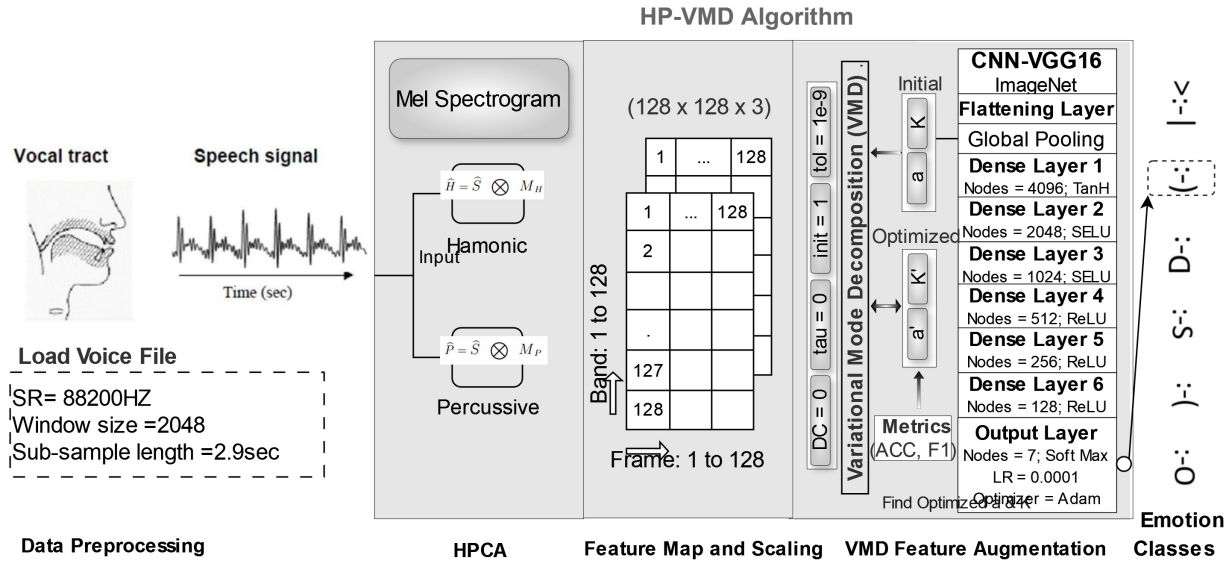

Fig. 1. Proposed HP-VMD framework to integrate HPC with VMD for enhanced SER

experiments assessing the proposed HP-VMD algorithm across different VMD hyperparameters and presented in the result section.

A. Results

Several metrics were employed to evaluate the proposed HP-VMD model efficacy, including F1 score, test set accuracy, and confusion matrix. Comparing the proposed HP-VMD approach for SER with a baseline model, operating under a similar framework but without integrating HPCA and VMD, confirmed HP-VMD superiority. The comparative analysis altered sample rate, window size, K and α.

Figure 2 shows that the proposed HP-VMD algorithm efficient functionality is particularly evident in processing Mel spectrograms and its harmonic and percussive components. This is most noticeable in Fig. 2(c), with considerably improved distinction for distinct frequency energy magnitude compared with baseline models (Figs. 2(a) and (b)). The HP-VMD algorithm enhances frequency component resolution, providing clearer and more pronounced differentiation among various frequency bands. This improvement is crucial to accurately capture nuances of emotional expressions in speech, ensuring that subtle variations in tone and intensity are more effectively detected and classified.

Table II shows the confusion matrix for the proposed and comparison baseline models. The proposed model achieves considerably improved proficiency, recognizing anger, fear, disgust, and sadness emotions with high accuracy (ACC = 98.01%, 98.04%, 98.41%, and 100%, respectively). However, it exhibits modest challenges to accurately predict neutral and boredom emotions (ACC = 93.24% and 88.97% accuracy, respectively). The proposed approach achieves average peak accuracy = 96.67% on the EMO-DB dataset. Additional insights regarding model implementation utilizing Python's Keras framework, and extended experimental results and visualizations can be accessed at our GitHub repositories[1].

TABLE I
EMPIRICAL FINDINGS (%) FOR F1 SCORE (F1) AND CLASSIFICATION TEST ACCURACY (ACC), ARE DISPLAYED FOR DIFFERENT COMBINATIONS OF DECOMPOSITION PARAMETERS K AND α, USING VARIOUS ACOUSTIC FEATURE EXTRACTION TECHNIQUES.

Features:	Feature-based Model Performance Analysis							
	Databases	α=2000, K=4		α=2000, K=6		α=3000, K=6		α=4000, K=6	
	EMO/RAV	Acc	F1	Acc	F1	Acc	F1	Acc	F1
SP-MS	EMODB	89.55	90.36	88.64	89.55	88.11	88.95	**95.11**	96.11
	RAVDESS	68.23	68.55	64.73	64.96	68.21	68.92	61.81	61.79
SP-MF	EMODB	58.84	58.86	66.15	66.07	65.19	65.07	**67.34**	67.98
	RAVDESS	64.21	64.69	61.36	61.55	65.28	65.95	64.19	64.68
CH-MF	EMODB	53.1	53.92	56.16	56.42	**60.87**	60.18	58.12	58.57
	RAVDESS	42.64	41.77	53.29	52.14	55.61	56.80	51.81	51.44
M-C-M	EMODB	86.27	86.11	87.01	87.95	**93.09**	93.07	92.44	92.37
	RAVDESS	58.25	59.11	59.48	59.21	52.28	52.88	51.70	51.10
MS	EMODB	91.93	91.11	93.25	93.89	93.92	93.91	**95.84**	95.12
	RAVDESS	64.21	64.26	61.04	61.67	65.07	65.12	64.06	64.12
HPCA	EMODB	88.62	89.85	89.76	89.08	89.2	89.13	**91.92**	91.11
	RAVDESS	68.33	68.12	63.37	63.79	63.57	63.78	62.38	62.42
MS-VMD	EMODB	90.1	90.2	91.91	91.98	**95.06**	96.01	94.54	94.13
	RAVDESS	64.08	64.12	66.25	66.68	69.28	69.94	68.21	68.14
HP-VMD	EMODB	95.21	95.2	94.35	94.36	**96.67**	96.63	95.41	95.52
	RAVDESS	65.29	65.35	64.25	64.89	65.65	65.66	67.13	67.12

Abbreviations: M-C-M: 3D Mel spectrogram + chromagram + MFCCs; SP-MS: spectral + 2D Mel spectrogram + spectral; CH: chromagram; MF: MFCC; TZ: 1D-Tonnetz; MS: Mel spectrogram; HPCA: harmonic-percussive components analysis; VMD: variational mode decomposition; HP-VMD: proposed model
The best results in different feature concatenation settings are highlighted in bold.

Table III compares the proposed HP-VMD model with recent state-of-the-art outcomes. The HP-VMD model surpasses all previous models, achieving superior test accuracy and establishing a new benchmark for performance.

V. CONCLUSION

This study proposed integrating harmonic-percussive component analysis (HPCA) and variational mode decomposi-

[1]https://github.com/DavidHason/hp-vmd

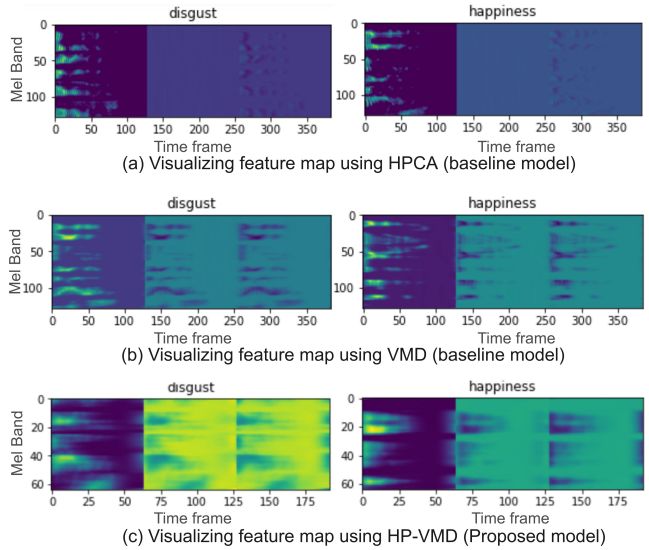

Fig. 2. Proposed HP-VMD algorithm improved efficiency processing the Mel spectrogram acoustic feature (c) provides greater distinction between frequency energy magnitudes compared with other baseline models (a) and (b).

TABLE II
CONFUSION MATRIX (%) SHOWING THE TEST ACCURACY IN 7 DIFFERENT CLASSES ON THE EMO-DB DATASET WITH AN AVERAGE TEST ACCURACY OF 96.67% .

Emotion:	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	98.01	0	0	0	1.99	0	0
Boredom	0	88.97	0	0	0	6.75	4.28
Disgust	0	0	98.41	0	1.59	0	0
Fear	0	0	0	98.04	0	0	1.96
Happiness	0	0	0	0	100	0	0
Neutral	0	1.88	0	0	0	93.24	4.88
Sadness	0	0	0	0	0	0	100

TABLE III
PROPOSED AND VARIOUS PREVIOUS STATE-OF-THE-ART MODELS ON EMODB AND RAVDESS DATABASES

Method proposed by	Feature extraction	Learning	Accuracy (%)
Dendukuri et al. [14]	Statistic-MF-SP	SVM-VMD	61.2
Hajarol. et al. [15]	MS-MF	CNN	72.21
Wang et al. [16]	FT-MFCCs	SVM	73.3
Kown et al. [17]	Spectrogram	Deep SCNN	79.50
Badsha et al. [5]	Spectrogram	CNN	80.79
Huang et al. [6]	Spectrogram	CNN	85.2
Issa et al. [18]	MF-CH-MS-CT-TZ	VGG16	86.10
Meng et al. [19]	Delta(log MS)	CNN-LSTM	90.78
Rudd et al. [21]	HPCA	VGG16-MLP	92.79
Demircan et al. [22]	LPC+MFCCs	SVM	92.86
Zhao et al. [4]	log MS	CNN-LSTM	95.89
Rudd et al. [27]	MS-CH-MF	VMD-VGG16	96.09
Proposed	HPC-MS	**HP-VMD + VGG16**	**96.67**

Abbreviations: MF-CH-MS-CT-TZ: MFCC + chromagram + Mel spectrogram + contrast + Tonnetz; MS-MF: Mel spectrogram + MFCC; log MS: log Mel spectrogram; Statistic-MF-SP: 45d mode statistical + MFCC + spectral; MSF: modulation spectral features; FT-MFCCs: Fourier parameter + MFCC; SP-MS: spectral + 2D Mel spectrogram + spectral; CH: chromagram; MF: MFCC; TZ: 1D Tonnetz; MS: Mel spectrogram; MS-CH-MF: 3D Mel spectrogram + chromagram + MFCC; HPCA: harmonic-percussive component analysis +log Mel spectrogram; VMD: variational mode decomposition; HP-VMD: proposed model;
Best results for both databases are indicated in bold font

tion (VMD) for speech emotion recognition and successfully demonstrated the HP-VMD algorithm's effectiveness to enhance SER. The HP-VMD algorithm enhances feature extraction precision and optimizes decomposition, preventing information loss from mode duplication and mixing. By employing VMD as a dynamic harmonic and percussive acoustic feature augmentation method that inputs enhanced feature vectors into the VGG16 CNN model, this approach achieves superior classification accuracy, with a notable 96.67% on benchmarks like the Berlin EMO-DB database.

Future studies research could expand the model capabilities to include real-time processing and broader emotional and dialectical ranges, potentially enhancing applications in interactive systems and health monitoring. Thus, the proposed HP-VMD model provides a significant advancement in affective computing, setting a new standard for empathetic human computer interaction.

ACKNOWLEDGMENT

This work is partially supported by the Australian Research Council under grant number: DP22010371, LE220100078, DP200101374 and LP170100891

REFERENCES

[1] Alshamsi, H., Kepuska, V., Alshamsi, H., Meng, H. (2018, November). Automated facial expression and speech emotion recognition app development on smart phones using cloud computing. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 730–738). IEEE.

[2] Carvalho, V. R., Moraes, M. F., Braga, A. P., Mendes, E. M. (2020). Evaluating five different adaptive decomposition methods for EEG signal seizure detection and classification. Biomedical Signal Processing and Control, 62, 102073.

[3] Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., ... Baik, S. W. (2019). Deep features-based speech emotion recognition for smart affective services. Multimedia Tools and Applications, 78, 5571–5589.

[4] Zhao, J., Mao, X., Chen, L. (2019). Speech emotion recognition using deep 1D 2D CNN LSTM networks. Biomedical signal processing and control, 47, 312–323.

[5] Rozgic, V., Ananthakrishnan, S., Saleem, S., Kumar, R., Prasad, R. (2012, December). Ensemble of SVM trees for multimodal emotion recognition. In Proceedings of the 2012 Asia Pacific signal and information processing association annual summit and conference (pp. 1–4). IEEE.

[6] Huang, Z., Dong, M., Mao, Q., Zhan, Y. (2014, November). Speech emotion recognition using CNN. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 801–804).

[7] Fitzgerald, D.: Harmonic/percussive separation using median filtering. In: Proceedings of the International Conference on Digital Audio Effects (DAFx). vol. 13, pp. 1–4 (2010)

[8] Dragomiretskiy, K., Zosso, D. (2013). Variational mode decomposition. IEEE transactions on signal processing, 62(3), 531–544.

[9] Lal, G. J., Gopalakrishnan, E. A., Govind, D. (2018). Epoch estimation from emotional speech signals using variational mode decomposition. Circuits, Systems, and Signal Processing, 37, 3245–3274.

[10] Zhang, M., Hu, B., Zheng, X., Li, T. (2020, December). A novel multidimensional feature extraction method based on VM and WPD for emotion recognition. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1216–1220). IEEE.

[11] Dennis, J., Tran, H. D., Li, H. (2010). Spectrogram image feature for sound event classification in mismatched conditions. IEEE signal processing letters, 18(2), 130–133.

[12] Shrawankar, U., Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: A comparative study. arXiv preprint arXiv:1305.1145.

[13] Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. Proceedings of the IEEE, 66(1), 51–83.

[14] Dendukuri, L. S., Hussain, S. J. (2022). Emotional speech analysis and classification using variational mode decomposition. International Journal of Speech Technology, 25(2), 457–469.

[15] Hajarolasvadi, N., Demirel, H. (2019). 3D CNN-based speech emotion recognition using k-means clustering and spectrograms. Entropy, 21(5), 479.

[16] Wang, K., An, N., Li, B. N., Zhang, Y., Li, L. (2015). Speech emotion recognition using Fourier parameters. IEEE Transactions on affective computing, 6(1), 69–75.

[17] Mustaqeem, Kwon, S. (2019). A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors, 20(1), 183.

[18] Issa, D., Demirci, M. F., Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894.

[19] Meng, H., Yan, T., Yuan, F., Wei, H. (2019). Speech emotion recognition from 3D log-Mel spectrograms with deep learning network. IEEE access, 7, 125868–125881.

[20] Honkola, T., Ruokolainen, K., Syrjänen, K.J.J. et al. Evolution within a language: environmental differences contribute to divergence of dialect groups. BMC Evol Biol 18, 132 (2018). https://doi.org/10.1186/s12862-018-1238-6

[21] Rudd, D. H., Huo, H., Xu, G. (2022, May). Leveraged Mel spectrograms using harmonic and percussive components in speech emotion recognition. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 392–404). Cham: Springer International Publishing.

[22] Demircan, S., Kahramanli, H. (2018). Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech. Neural Computing and Applications, 29, 59–66.

[23] Hason Rudd, D., Huo, H., Xu, G. (2023, May). An extended variational mode decomposition algorithm developed speech emotion recognition performance. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 219–231). Cham: Springer Nature Switzerland.

[24] Hestenes, M. R. (1969). Multiplier and gradient methods. Journal of Optimization Theory and Applications, 4(5), 303–320.

[25] Pandey, P., Seeja, K. R. (2022). Subject independent emotion recognition from EEG using VMD and deep learning. Journal of King Saud University-Computer and Information Sciences, 34(5), 1730-1738.

[26] Zhang, M., Hu, B., Zheng, X., Li, T. (2020, December). A novel multidimensional feature extraction method based on vmd and wpd for emotion recognition. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1216-1220). IEEE.

[27] Hason Rudd, D., Huo, H., Xu, G. (2023, May). An extended variational mode decomposition algorithm developed speech emotion recognition performance. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 219-231). Cham: Springer Nature Switzerland.

[28] Taran, S. (2023). A nonlinear feature extraction approach for speech emotion recognition using VMD and TKEO. Applied Acoustics, 214, 109667.

[29] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B. (2005, September). A database of German emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).

[30] Livingstone, S. R., Russo, F. A. (2018). 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, 13(5), e0196391.

[31] Mishra, S. P., Warule, P., Deb, S. (2023). Variational mode decomposition based acoustic and entropy features for speech emotion recognition. Applied Acoustics, 212, 109578.