An Extended Variational Mode Decomposition Algorithm Developed Speech Emotion Recognition Performance

David Hason Rudd[1], Huan Huo[1(✉)], and Guandong Xu[1,2(✉)]

[1] The University of Technology Sydney, 15 Broadway, Ultimo, Australia
{david.hasonrudd,huan.huo,guandong.xu}@uts.edu.au
[2] Data Science Institute, 15 Broadway, Ultimo, Australia

Abstract. Emotion recognition (ER) from speech signals is a robust approach since it cannot be imitated like facial expression or text based sentiment analysis. Valuable information underlying the emotions are significant for human-computer interactions enabling intelligent machines to interact with sensitivity in the real world. Previous ER studies through speech signal processing have focused exclusively on associations between different signal mode decomposition methods and hidden informative features. However, improper decomposition parameter selections lead to informative signal component losses due to mode duplicating and mixing. In contrast, the current study proposes VGG-optiVMD, an empowered variational mode decomposition algorithm, to distinguish meaningful speech features and automatically select the number of decomposed modes and optimum balancing parameter for the data fidelity constraint by assessing their effects on the VGG16 flattening output layer. Various feature vectors were employed to train the VGG16 network on different databases and assess VGG-optiVMD reproducibility and reliability. One, two, and three-dimensional feature vectors were constructed by concatenating Mel-frequency cepstral coefficients, Chromagram, Mel spectrograms, Tonnetz diagrams, and spectral centroids. Results confirmed a synergistic relationship between the fine-tuning of the signal sample rate and decomposition parameters with classification accuracy, achieving state-of-the-art 96.09% accuracy in predicting seven emotions on the Berlin EMO-DB database.

Keywords: Speech emotion recognition (SER) · Variational mode decomposition (VMD) · Sound signal processing · Convolutional neural network (CNN) · Acoustic features

1 Introduction

Word meaning is often conveyed by the tone of voice, although human emotions are not solely conveyed through the words used, but also through by modifying facial expressions and vocal tone. Thus, changing voice characteristics is how

H. Kashima et al. (Eds.): PAKDD 2023, LNAI 13937, pp. 219–231, 2023.
https://doi.org/10.1007/978-3-031-33380-4_17

most humans express different emotions [25]. Consequently, considerable human-computer interaction research has analyzed speech signal emotion recognition (ER) where using other popular semantic analysis methods like wav2vec2.0 [5] are not trustworthy. Several applications employed variational mode decomposition (VMD) [11] in different fields such as medical science, structural engineering, and sound engineering [2,17,23]. Signal based ER employs various instantaneous signals, including electrodermal activity, blood volume pulse, galvanic skin response, electrocardiogram (ECG), Electroencephalography (EEG), and speech, are commonly categorized into several decomposed modes due to the complexity and nonstationary nature of them, which allows latent factors and patterns to be extracted more easily. Nonstationary signal properties and its components make mean short time Fourier transform (STFTs) are not always suitable, and previous studies have mostly considered these approaches in isolation [8]. VMD decomposes signals into modes with a narrowband around a center frequency; it can overcome STFT limitation and EMD mode mixing effects. Therefore, we were motivated to apply VMD for speech signal processing.

Acoustic feature selection is essential for SER to describe various voice signal aspects captured from different features [6]. Acoustic features include time-frequency, time, and frequency domain representations. Extracted features from time-frequency domains carry more informative data than the other domains, and better capture latent emotion content from speech signals [28]. Several previous studies used VMD method to analyze signals, extracting features from the decomposed signals. However, we proposed VGG-optiVMD, utilizing a VMD based feature augmentation method to enrich predictors and maximize emotion classification accuracy. Results from the proposed VGG-optiVMD approach on several common publicly available databases confirm significant ER improvement compared with previous approaches. The main contributions from this study can be summarized as follows.

- To our best knowledge, this study is the first to employ VMD as a dynamic acoustic feature augmentation method for SER performance.
- The proposed VGG-optiVMD algorithm automatically selects optimum decomposition parameters for VMD.
- A robust classification accuracy was achieved with a state-of-art result 96.09%.

2 Related Works

Dendukuri et al. [10] decomposed the speech signal into three components sampling 16000 Hz over 20 ms frames, then input various mode central frequency statistical parameters to a support vector machine (SVM) classifier. Lal et al. [20] empirically demonstrated VMD advantages to decompose speech signals in the correct central frequency and subsequently estimated epoch locations from noise degraded emotional speech signal. Zhang et al. [33] proposed multidimensional feature extraction for EEG signal emotion recognition combining wavelet packet decomposition (WPD) with VMD to break down an EEG signals and extract

wavelet packet entropy, modified multiscale sample entropy, fractal dimension, and first difference of each emotional variational mode functions as feature components. They subsequently demonstrated robust results using a random forest (RF) classifier on the DEAP dataset [18]. Khare et al. [17] reduced reconstruction error using meta-heuristic techniques to condensing from 16 to 1 dimension using eigenvector centrality method channel selection on EEG signals. They subsequently improved Optimized variational mode decomposition (O-VMD) accuracy by 5% compared with traditional VMD on the dataset of four emotions that built by themselves.

Pandey [24] proposed subject-independent emotion recognition using VMD and deep neural networks (VMD-DNN) on the benchmark DEAP dataset. Two features, first difference and power-spectral-density used since were sufficient to recognize calm, happy, sad, and angry emotions. SVM and DNN classifier accuracy was improved by employing VMD based feature extraction.

Several previous studies considered STFT signal decomposition techniques for SER. Few previous studies employed VMD to analyze speech signals mainly processing EEG signals through VMD for ER. To the best of our knowledge, the current study is the first to employ VMD to enrich multidimensional feature vectors to enhance VGG-16 network learning.

3 Proposed Methodology

The main aim for decomposition-based speech signal processing via VMD method is to constrain noise and interference frequencies to enhance signal data decoding.

3.1 Variational Mode Decomposition

Variational mode decomposition is a popular technique for decomposing non-stationary signals into sub-signals or modes, where mode contains a specific meaningful property from the original signal in a narrow bandwidth around the center frequency. The VMD adaptive algorithm reduces the original signal complexity [11]. The VMD algorithm applies the Wiener filter, Hilbert transform, analytical signals, and frequency mixing. The two main VMD objects are to constrain the bandwidth for each IMF center frequency and reconstruct the original signal from the sum of all modes. First, the Hilbert transform filters frequencies on the negative side of the spectrum, and then shifts the obtained bandwidth to the modes central frequency. Second, the obtained spectrum is shifted to the baseband region via a modulator function to obtain bandwidth around central frequency ω. Finally, H1 Gaussian smoothness for the demodulation signal is used to estimate the bandwidth. Thus, constraining the L^2 norm squared gradient [11] defines the optimization problem (1),

$$\begin{aligned} \min_{\{g_k\},\{\omega_k\}} \left\{ \sum_{k=1}^{K} \left\| \frac{\partial}{\partial_t} \left[\left(\delta(t) + \frac{j}{\pi t}\right) * g_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}, \\ \text{subject to:} \quad \sum_{k=1}^{K} g_k(t) = g(t), \end{aligned} \tag{1}$$

where the partial derivative $\frac{\partial}{\partial_t}[.]$ minimizes variation in the obtained bandwidth; $g(t)$ is the original speech signal frame; $g_k(t)$ is the kth mode for $g(t)$; K is the total number of modes; $\omega_k = \{w1, \ldots, wk\}$ is the mode center frequency, and a convenient way to reference the center frequencies for the set of K modes; $e^{-j\omega_k t}$ is a modulator function to shift the spectrum for each mode to the baseband.

The analytical signal generated by applying the Hilbert transform $\frac{j}{\pi t}$ and unit impulse function $\delta(t)$ as shown in equation (1). The $\delta(t)$ denotes to the Dirac delta distribution known as a unit impulse so that its value is zero everywhere and infinite at original signal. The original voice signal can be reproduced by solving the constraint optimization (1), which can be simplified using an augmented Lagrangian multiplier to transform it into an unconstrained problem (2),

$$\begin{aligned}\mathcal{L}\left(g_k, \omega_k, \lambda\right) := \alpha \sum_{k=1}^{K} \left\| \frac{\partial}{\partial_t}\left[\left(\left(\delta(t) + \frac{j}{\pi t}\right) * g_k(t)\right) e^{-j\omega_k t}\right]\right\|^2 \\ + \left\| g(t) - \sum_{k=1}^{K} g_k(t) \right\|_2^2 + \left\langle \lambda(t), g(t) - \sum_{k=1}^{K} g_k(t) \right\rangle,\end{aligned} \tag{2}$$

where, λ is a time-dependent Lagrangian multiplier, and α is a bandwidth control parameter. The unconstrained Lagrangian problem (2) can be solved to obtain the frequency and the modes using the alternate direction method of multipliers (ADMM) [11,14,27] optimization in spectral domain. However, optimization outcomes are the same for the frequency and time domains. Hence, mode $g_k(\omega)$ can be updated in the spectral domain,

$$\hat{g}_k^{n+1}(\omega) \leftarrow \frac{\hat{g}(\omega) - \sum_{i<k} \hat{g}_i^{n+1}(\omega) - \sum_{i>k} \hat{g}_i^{n}(\omega) + \frac{\hat{\lambda}^n(\omega)}{2}}{1 + 2\alpha\left(\omega - \omega_k^n\right)^2}. \tag{3}$$

Updating is obtained using the Wiener filter for the current residual using the signal prior $1/(\omega - \omega_k)^2$ to restrain variation across the central frequency minimum, providing the updated mode center frequency ω_k as

$$\hat{\omega}_k^{n+1} = \frac{\int_0^\infty \omega \left|\hat{G}_k(\omega)\right|^2 \mathrm{d}\omega}{\int_0^\infty \left|\hat{G}_k(\omega)\right|^2 \mathrm{d}\omega} \tag{4}$$

where $\hat{G}_k(\omega)$ is the Fourier transformed for $g_k^{n+1}(t)$. A better decomposed signal can be obtained by reconstructing the original signal as the sum of modes and estimating bandwidth using the Wiener filter. Details of the VMD algorithm are provided in [11]. To leverage VMD effectiveness, we proposed the VGG-optiVMD algorithm for automatically selecting optimum α and K by analyzing different decomposition parameter effects on classification accuracy.

3.2 Proposed VGG-optiVMD

Reconstruction error for a decomposed signal can be reduced by selecting optimum K and α. Improper decomposition parameter selection will create duplicate modes, causing signal information losses consequently reduced classifier performance.

Algorithm 1. Proposed VGG-optiVMD algorithm

Input: $g(t)$ is a preprocessed speech signal converted to feature vectors.
Output: Decomposes of signal $g(t)$ and Optimum value of α and K
Initialization: The value of modes K and α;
the tolerance of convergence criterion τ; $\left\{\hat{g}_k^1\right\}, \left\{\hat{\omega}_k^1\right\}, \hat{\lambda}^1$; $n = 0$
Repeat:
1: $n = n + 1$,
2: **for** k=1 : K **do**
3: update $\hat{g}_k$ for all $\omega \geq 0$ by Eq. (3) and ω_k by Eq. (4)
4: **end for**
5: Upgrade the Lagrangian multiplier λ for the dual accent $\forall\omega 0$:

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

Until:
6: convergence: $\sum_{k=1}^{K} \|\hat{g}_k^{n+1} - \hat{g}_k^n\|_2^2 / \|\hat{g}_k^n\|_2^2 < \in$.
7: **return** $\{g_1(t), g_2(t), \ldots, g_K(t)\}$= IMFs; subtract of all sub signals
8: Set Parameters τ=0; DC=0; init=1; tol=1e-9; K=2; α=2000
9: Decompose signal $g(t)$
10: Record training set accuracy, and F1 score in VGG16 classifier.
11: **while** max(ACC) do
12: **if** ACC==max; $\alpha \leq 6000$; $K \leq 8$ **then**
13: The optimum value of K and α is obtained.
14: **else**
$K = K + 1$; α=α+1000 go to step 9
15: **end if**
16: **end while**
17: Identify optimum value of decomposition parameters α and K while tol=1e-9, DC=0, init=1, and τ=0

One drawback for VMD is that finding decomposition parameters K and α to provide optimum performance challenging. In contrast, in our method we automate optimum VMD decomposition parameter selection using a feedback loop from the VGG16 flattening output layer. Algorithm 1 shows the proposed optimized VMD algorithm (VGG-optiVMD). The key strength for VGG-optiVMD is generality and reproducibility across different databases for real-world multimedia applications, e.g., ER for customer satisfaction analysis.

3.3 Feature Extraction, Data Augmentation, and Classification

Essential and informative acoustic features in the time-frequency domain include the Mel spectrogram, chromograms, spectral contrasts, tonnetz, and Mel-frequency cepstral coefficients (MFCCs) [1,13] are extracted and subsequently employed in various combinations to generate multidimensional feature vectors. Figure 1 shows the proposed framework to train CNN-VGG16 [29] to extract enriched feature vectors and classify seven emotions: anger, boredom, happy, neutral, disgust, sadness, and fear on two databases EMODB [7] and RAVDESS [21].

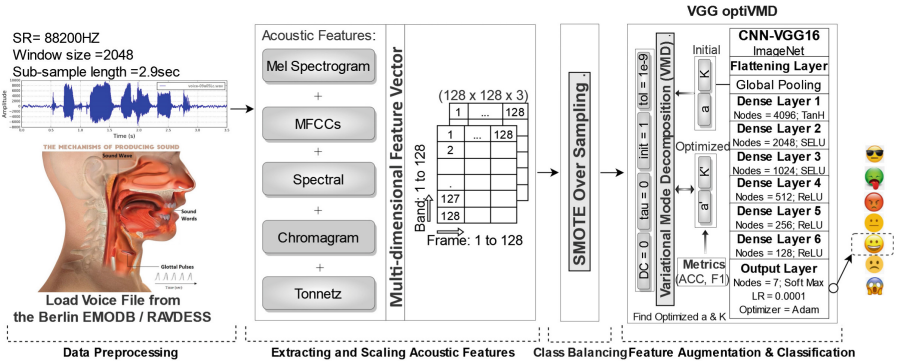


Fig. 1. Proposed model development workflow: extracted features are enriched using the VGG-optiVMD to automatically identify K and α.

Figure 1 shows the model development proceeds. First, the voice signal is sampled 88400 Hz and five well-known acoustic features extracted and reshaped into a single ($128 \times 128 \times 3$) feature vector and second the SMOTE [21] oversampling strategy is applied to compensate for minority classes and reduce model bias. Furthermore, the testing and training features are randomly partitioned into 20% and 80% sets, respectively. Subsequently, the proposed VGG-optiVMD algorithm is applied to decode frequency statistical properties at specific times that distinguish emotions within the feature vector. Finally, the VGG network is trained on the augmented feature vector to classify emotions into seven classes.

4 Experiment Setup

Several experiments were performed on nine different feature vectors to identify the proposed VGG-optiVMD algorithm effectiveness using. The details of network implementations are available in our GitHub repository[1].

4.1 Modelling

The aim of modeling was to enhance informative data within the feature vectors and avoid overfitting. Augmentation effects on classification accuracy were assessed using diverse K and α sets. Optimal K and α was assessed iteratively until robust classification accuracy was achieved or the break loop condition reached. K and α were set to a wide range of 3–8 and 1000–6000, respectively, based on empirical experiments since there was no significant improvement in prediction accuracy outside those ranges. The VGG16 architecture used the ADAM optimizer with learning rate = 0.0001; six fully connected hidden layers with ReLU, SELU, and TanH activation functions; epochs = 50, batch size = 4; and SoftMax function for the output layer.

[1] https://github.com/DavidHason/VGG-optiVMD.

5 Result and Discussion

To assess the effectiveness of our VMD-based feature augmentation method several evaluation metrics were employed including F1 score, training set accuracy, and confusion matrix. Analyzing the results of the baseline model, which is built with the same framework simply without VMD-based feature vector augmentation, helps us to justify the power of the VGG-optiVMD in SER. Therefore, we attempted to evaluate the model performance through variation of sample rate, window size, K and α without using VMD (baseline model) and with VMD (proposed model). As shown in Fig. 2, unlike the baseline model, the proposed model performed better with a larger sampling rate and window size. Moreover, the highest train set accuracy and F1 score were obtained via VGG-optiVMD, proving that our VMD-based feature augmentation method significantly improved the classification accuracy.

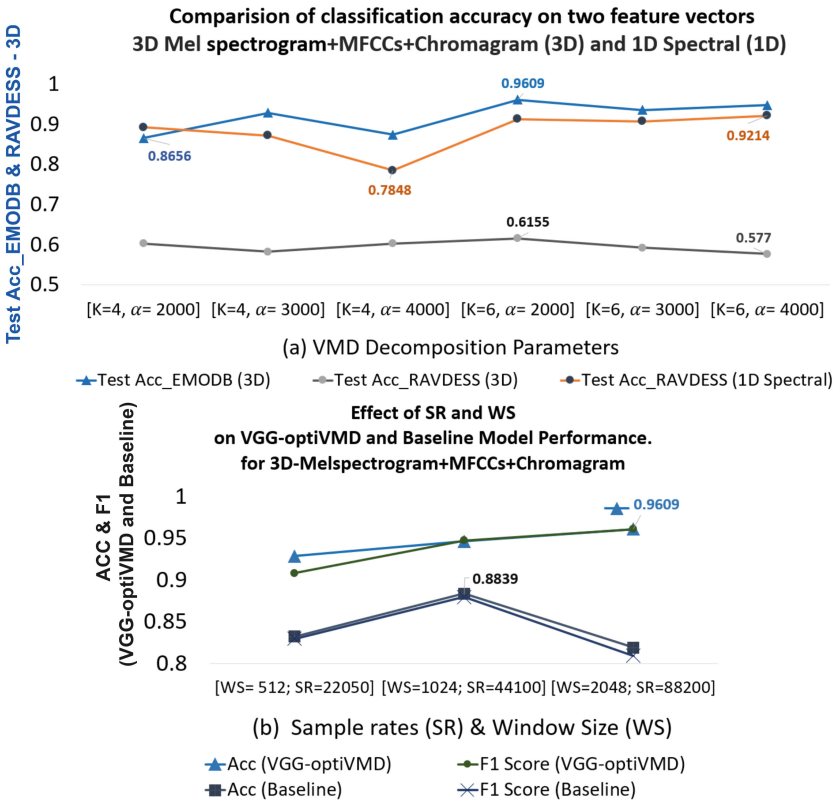


Fig. 2. The model performance is assessed by different signal sampling rates and VMD parameters K and α. Graph (a) The VGG-optiVMD identified the set of $K = 6$ and $\alpha = 2000$ as optimum value. Graph(b) represents the effect of various ranges of sample rate and window size on the proposed and baseline model in EMODB. The highest accuracy can be achieved by SR = 88200 and WS = 2048.

Based on the experiment results shown in Table 1, there is a correlation between the number of modes K, bandwidth control parameter α and classification accuracy. The different acoustic features are enriched with various sets of decomposition parameters. Results indicated that higher accuracy was obtained for K (4–6) and α (2000– 4000) in both datasets, although VGG-optiVMD is set to a limited range of α (1000–10000) and K (2–8) due to increasing a heavy computational load when K value is over 8 with sample rate 88400. This limitation can be considered a functional constraint of VGG-optiVMD. Nevertheless, a state-of-the-art result was achieved with the accuracy of 96.09% with K=6 and α=2000 as demonstrated in Table 1. The Fig. 3 shows the efficient functionality of VGG-optiVMD on the feature vector 3D-Mel Spectrogram+MFCCs+Chromagram. Figure (a) represents the feature before applying VMD based data augmentation, and figure (b) clearly shows that the informative frequencies are distinguished on the feature vector by acquiring higher distinction energies represented in time-frequency domain after applying the data augmentation method. Therefore, the implications of this finding can improve

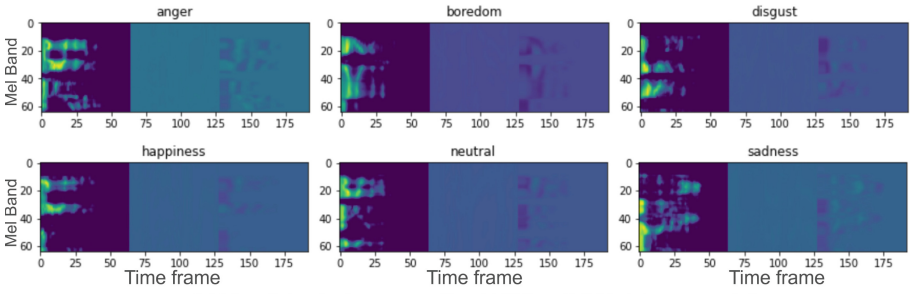
Table 1. Empirical results (%) of emotion classification accuracy (ACC) and F1-score (F1) are demonstrated through different sets of decomposition parameters α and K, that are selected automatically by the VGG-optiVMD algorithm.

Features:	VMD Decomposition Parameters										
	Databases	K=4, α=2000		K=4, α=4000		K=6, α=2000		K=6, α=3000		K=6, α=4000	
	EMO/RAV	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CH	EMODB	68.54	68.37	81.63	81.47	94.05	94.88	94.90	91.10	95.41	95.11
	RAVDESS	70.23	70.55	82.73	82.96	85.21	85.92	79.81	79.79	47.49	46.53
MS	EMODB	91.84	91.86	93.15	93.07	95.19	95.07	95.34	94.98	95.92	94.89
	RAVDESS	64.21	64.69	71.36	71.55	75.28	75.95	84.19	84.68	87.25	88.11
MF	EMODB	48.1	46.92	65.16	64.42	64.87	65.18	56.12	56.57	67.64	66.9
	RAVDESS	42.64	41.77	53.29	52.14	55.61	56.80	51.81	51.44	41.86	40.46
SP	EMODB	94.27	93.11	93.01	92.95	93.88	93.07	93.44	93.37	94.02	93.87
	RAVDESS	89.25	90.11	78.48	79.21	91.28	92.88	90.70	90.10	**92.14**	93.55
TZ	EMODB	74.93	75.11	91.25	90.89	88.92	88.91	91.84	91.12	92.44	92.10
	RAVDESS	48.21	48.26	51.04	51.67	52.07	52.12	49.06	49.12	51.98	52.23
MS+SP	EMODB	89.62	90.85	88.76	89.08	88.2	88.13	95.92	96.11	95.41	95.12
	RAVDESS	78.33	78.12	74.37	74.79	78.52	78.78	81.38	81.42	81.84	81.91
MF+SP	EMODB	58.1	58.2	66.91	66.98	65.16	65.11	62.54	62.13	67.64	67.21
	RAVDESS	53.08	53.12	56.25	56.68	60.28	60.94	58.21	58.14	54.7	54.06
MF+CH	EMODB	85.21	85.2	84.35	84.36	90.14	90.13	87.41	87.52	90.82	90.82
	RAVDESS	51.29	51.35	54.25	54.89	53.65	54.66	55.13	55.12	56.08	56.84
M+M+C	EMODB	86.56	86.42	87.41	87.35	**96.09**	96.04	93.54	93.42	94.73	95.98
	RAVDESS	60.28	60.11	60.28	60.84	61.55	62.36	59.25	60.87	57.70	57.56

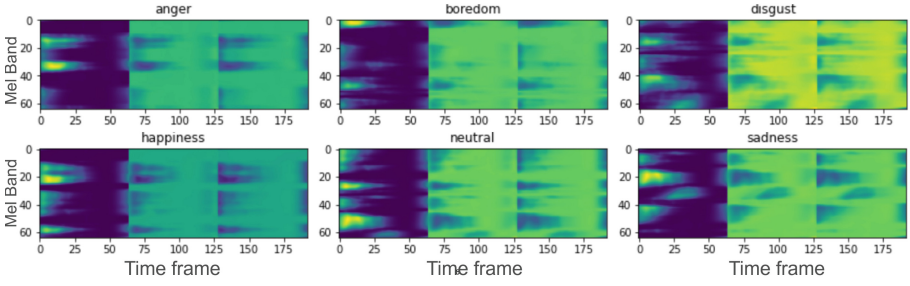
Features abbreviation: M+M+C: 3D-Mel Spectrogram+MFCCs+Chromagram; MS+SP: 2D-Mel Spectrogram+Spectral; CH: Chromagram; MF: MFCCs; TZ: 1D-Tonnetz;
The best results on both databases are indicated in bold font.

Table 2. Visualization of the model performance with confusion matrix (%) for the 3D-Mel Spectrogram+MFCCs+Chromagram with test accuracy = %96.09 on the Berlin EMO-DB dataset.

Emotion:	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	**95.24**	0	0	0	4.76	0	0
Boredom	0	**95.24**	0	0	0	0	4.76
Disgust	0	0	**100.00**	0	8	0	0
Fear	0	0	0	**94.05**	0	0	0
Happiness	8.33	0	0	0	**91.67**	0	0
Neutral	0	2.38	0	0	1.19	**96.43**	0
Sadness	0	0	0	0	0	0	**100**



(a) Visualizing feature map without VGG-optiVMD data augmentation



(b) Visualizing feature map with VGG-optiVMD data augmentation

Fig. 3. The efficient functionality of VGG-optiVMD on the feature vector 3D-Mel Spectrogram+MFCCs+Chromagram clearly shows a higher distinction in the energy magnitude of frequencies in (b).

the learning process in VGG16 and result in better prediction accuracy. The confusion matrix in Table 2 demonstrates the high performance of the classification model with accuracy above 90% for all classes. Nevertheless, the model performs poorly when predicting happiness and anger emotions due to the similarity of signal attributes such as intensity, frequency and harmonic structure. The VGG-optiVMD method is compared with the most recent works, shown in Table 3, that our method outperforms previous models and achieves a state-of-the-art result

Table 3. Comparison of the proposed method with previous works on the EMODB and RAVDESS databases.

Method proposed by	Feature extraction strategy	Learning Net.	Acc(%)
Badshah et al. [3]	log Mel spectrogram	CNN	52
Dendukuri et al. [10]	45d- Mode statistical+MFCCs+Spectral	SVM-VMD	61.2
Zamil et al. [32]	13 MFCCs	Tree Model	70
Popova et al. [26]	Mel spectrograms	VGG16	71
Hajarol. et al. [12]	Mel spectrograms+MFCCs	CNN	72.21
Wang et al. [30]	Fourier Parameter+MFCCs	SVM	73.3
Kown et al. [19]	Spectrogram	Deep SCNN	79.50
Badsha et al. [4]	Spectrogram	CNN	80.79
Huang et al. [15]	Spectrogram	CNN	85.2
Issa et al. [16]	MFCCs+Chroma.+Mel spec.+Contrast+Tonnetz	VGG16	86.10
Meng et al. [22]	log Mel spec.+1st & 2nd delta(log Mel spec.)	CNN-LSTM	90.78
Wu et al. [31]	Modulation Spectral Features (MSFs)	SVM	91.60
Rudd et al. [28]	Harmonic-Percussive (HP)+log Mel spec	VGG16-MLP	92.79
Demircan et al. [9]	LPC+MFCCs	SVM	92.86
Zhao et al. [34]	log Mel spectrogram	CNN-LSTM	95.89
VGG-optiVMD	**3D-Mel spectrogram+MFCCs+Chromagram**	**VGG16-VMD**	**96.09**

in terms of accuracy. Moreover, the main advantage of the VGG-optiVMD is its generality, which can be employed independently for other acoustic features and different databases.

6 Conclusion

Speech signal processing is employed in some applications when we only have access to speech voice to detect emotions which is the first aim of this study, the second aim of this study is to introduce specific data augmentation techniques to enrich the extracted acoustic features by design of VGG-optiVMD, an extended VMD algorithm to improve SER performance.

The findings provide solid empirical confirmation of the key role of the sampling rate, the number of the decomposed mode, K and the balancing parameter of the data-fidelity constraint, α, in the performance of the emotion classifier. Taken together, these findings suggest that VMD decomposition parameters K (2–6) and α (2000–6000) are and EMODB databases. The proposed VGG-optiVMD algorithm improved the emotion classification to a state-of-the-art result with a test accuracy of 96.09% in the Berlin EMO-DB and 86.21% in the RAVDESS datasets. Further work needs to be done to establish whether extracting acoustic features only from informative decomposed modes can reduce computational load constraints. Therefore, the study should be repeated using the VMD algorithm before acoustic feature extraction process.

Acknowledgement. This work is partially supported by the Australian Research Council under grant number: DP22010371, LE220100078, DP200101374 and LP1701 00891

References

1. Aizawa, Kiyoharu, Nakamura, Yuichi, Satoh, Shin'ichi (eds.): PCM 2004. LNCS, vol. 3331. Springer, Heidelberg (2005). https://doi.org/10.1007/b104114
2. Alshamsi, H., Kepuska, V., Alshamsi, H., Meng, H.: 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 (2018)
3. Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W.: Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 International Conference on Platform Technology and Service (PlatCon), pp. 1–5 (2017)
4. Badshah, A.M., Rahim, N.: Ullah: Deep features-based speech emotion recognition for smart affective services. Multimedia Tools and Applications **78**(5), 5571–5589 (2019)
5. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460 (2020)
6. Basharirad, B., Moradhaseli, M.: Speech emotion recognition methods: A literature review. In: AIP Conference Proceedings, vol. 1891, p. 020105. AIP Publishing LLC (2017)
7. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., et al.: A database of german emotional speech. In: Interspeech. vol. 5, pp. 1517–1520 (2005)
8. Carvalho, V.R., Moraes, M.F., Braga, A.P., Mendes, E.M.: Evaluating five different adaptive decomposition methods for eeg signal seizure detection and classification. Biomed. Signal Process. Control **62**, 102073 (2020)
9. Demircan, S., Kahramanli, H.: Application of fuzzy c-means clustering algorithm to spectral features for emotion classification from speech. Neural Comput. Appl. **29**(8), 59–66 (2018)
10. Dendukuri, L.S., Hussain, S.J.: Emotional speech analysis and classification using variational mode decomposition. Int. J. Speech Technol, pp. 1–13 (2022)
11. Dragomiretskiy, K., Zosso, D.: Variational mode decomposition. IEEE Trans. Signal Process. **62**(3), 531–544 (2013)
12. Hajarolasvadi, N., Demirel, H.: 3d cnn-based speech emotion recognition using k-means clustering and spectrograms. Entropy **21**(5), 479–495 (2019)
13. Harte, C., Sandler, M., Gasser, M.: Detecting harmonic change in musical audio. In: Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia, pp. 21–26 (2006)
14. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**(5), 303–320 (1969)
15. Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using cnn. In: Proceedings of the 22nd ACM International Conference Media, pp. 801–804 (2014)
16. Issa, D., Demirci, M.F., Yazici, A.: Speech emotion recognition with deep convolutional neural networks. Biomed. Signal Process. Control **59**, 101894–101904 (2020)

17. Khare, S.K., Bajaj, V.: An evolutionary optimized variational mode decomposition for emotion recognition. IEEE Sens. J. **21**(2), 2035–2042 (2020)
18. Koelstra, S., Kolestra, S., et al.: Deap: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2011)
19. Kwon, S.: A cnn-assisted enhanced audio signal processing for speech emotion recognition. Sensors **20**(1), 183 (2019)
20. Lal, G.J., Gopalakrishnan, E., Govind, D.: Epoch estimation from emotional speech signals using variational mode decomposition. Circ. Syst. Signal Process. **37**(8), 3245–3274 (2018)
21. Livingstone, S.R., Russo, F.A.: 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 (2018)
22. Meng, H., Yan, T., Yuan, F., Wei, H.: Speech emotion recognition from 3d log-mel spectrograms with deep learning network. IEEE access **7**, 125868–125881 (2019)
23. Mousavi, M., Gandomi, A.H.: Structural health monitoring under environmental and operational variations using mcd prediction error. J. Sound Vib. **512**, 116370 (2021)
24. Pandey, P., Seeja, K.: Subject independent emotion recognition from eeg using vmd and deep learning. J. King Saud University-Comput. Inform. Sci. **34**(4), 1730–1738 (2019)
25. Pierre-Yves, O.: The production and recognition of emotions in speech: features and algorithms. Int. J. Hum Comput Stud. **59**(1–2), 157–183 (2003)
26. Popova, A.S., Rassadin, A.G., Ponomarenko, A.A.: Emotion recognition in sound. In: International Conference on Neuroinformatics, pp. 117–124 (2017)
27. Rockafellar, R.T.: A dual approach to solving nonlinear programming problems by unconstrained optimization. Math. Program. **5**(1), 354–373 (1973)
28. Rudd, D.H., Huo, H., Xu, G.: 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. Springer (2022). https://doi.org/10.1007/978-3-031-05936-0_31
29. Russakovsky, O., Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)
30. Wang, K., An, N., Li, B.N., Zhang, Y., Li, L.: Speech emotion recognition using fourier parameters. IEEE Trans. Affect. Comput. **6**(1), 69–75 (2015)
31. Wu, S., Falk, T.H., Chan, W.Y.: Automatic speech emotion recognition using modulation spectral features. Speech Commun. **53**(5), 768–785 (2011)
32. Zamil, A.A.A., Hasan, S., Baki, S.M.J., Adam, J.M., Zaman, I.: Emotion detection from speech signals using voting mechanism on classified frames. In: 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 281–285. IEEE (2019)
33. Zhang, M., Hu, B., Zheng, X., Li, T.: 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 (2020)
34. Zhao, J., Mao, X., Chen, L.: Speech emotion recognition using deep 1d & 2d cnn lstm networks. Biomed. Signal Process. Control **47**, 312–323 (2019)