SERBIAN JOURNAL OF ELECTRICAL ENGINEERING
Vol. 23, No. 3, October 2026, 523-538
UDC: 519.711:[159.942.3:811.21`282 **DOI: https://doi.org/10.2298/SJEE2603523M**
Original scientific paper

A Comprehensive Approach to Speech Emotion Recognition: Combining MFCC, Multifractal, Prosodic Features, and Volterra Filtering

Fathima Kunhi Mohamed[1], Abdul Haleem Pezumkattil Pottammal[2]

Abstract: This study proposes a new approach to speech emotion recognition by integrating traditional Mel-Frequency Cepstral Coefficients (MFCC) features with multifractal features derived from speech signals. Then a Sparse Adaptive Volterra Filter was used to improve these features. A Convolutional Neural Networks – Bidirectional Long Short- Term Memory (CNN-BiLSTM) was used to classify the emotions. Experiments were conducted on a self-built Malayalam emotional dataset of 3120 samples. Accuracy increased from 89.10% with the MFCC-only model to 93.43% when combined with multifractal features. This is an absolute improvement of 4.33%. Adding the Volterra filter raised the final accuracy to 95.51%, a total improvement of 6.41% over the MFCC-only baseline. Including pitch and formant features raised the accuracy further to 96.15%. For further validation, the method was also tested on the RAVDESS, a widely used benchmark dataset. Overall, the proposed framework gave high accuracy compared to baseline methods. These results show the potential of this methodology for practical emotion recognition applications, and points to the model's robustness and effectiveness in real-world scenarios.

Keywords: Malayalam, Sparse Adaptive Volterra filter, Emotion recognition.

1 Introduction

MFCCs are the most used spectral features in speech processing tasks, including Speech Emotion Recognition (SER). They effectively capture the perceptual characteristics of human auditory processing [1−3]. However, MFCCs primarily represent linear features. They may overlook complex emotional variations embedded in the nonlinear dynamics of speech. To overcome this

[1]University of Calicut, Farook College, Department of Computer Science, Kerala, India
fathimavk@gmail.com, https://orcid.org/0009-0004-8543-9115
[2]Kannur University, Department of Information Technology, Kerala, India
abdulhaleem@kannuruniv.ac.in, https://orcid.org/0000-0002-8172-8700

Colour versions of the one or more of the figures in this paper are available online at https://sjee.ftn.kg.ac.rs

© *Creative Common License CC BY-NC-ND*

limitation, researchers have increasingly explored nonlinear approaches like fractal and multifractal analysis to describe speech signal complexity [4, 5].

Multifractal features are derived from methods like Multifractal Detrended Fluctuation Analysis (MF-DFA) and Wavelet Transform Modulus Maxima (WTMM). These features effectively characterize the signal's scaling behavior and temporal irregularities [6, 7]. Reports have shown that combining multifractal features with MFCCs further improves classification performance in SER tasks using speech signals [8] and EEG signals [9]. Moreover, deep learning models such as CNNs and BiLSTM networks have shown to be effective in modeling both local patterns and temporal dependencies in speech signals [10, 11].

Prosodic features such as pitch, energy, and formant frequencies capture important emotional indicators. Sadness is often marked by lower pitch and reduced energy. Anger and happiness tend to show higher pitch and intensity. Prosodic features are informative, they are sensitive to speaker differences and recording conditions. This motivates the combination of complementary spectral and nonlinear features for more reliable emotion recognition.

Multifractal analysis describes the complex, nonlinear dynamics of speech signals which makes it useful for SER. Sarkar et al. employed Multifractal Detrended Fluctuation Analysis (MFDFA) and Wavelet Transform Modulus Maxima (WTMM) to extract features from emotional speech signals [12]. Their work showed that multifractal features are useful for SER. Mao and Chen introduced correlation dimension and fractal dimension as features for SER [13]. These measures capture the self-similarity and nonlinearity of speech signals and gave higher classification accuracy than traditional features with Hidden Markov Models and K-Nearest Neighbors. Tamulevičius et al. studied the use of fractal dimension-based features in a convolutional neural network (CNN) framework for cross-linguistic SER [14]. Lal et al. applied multifractal analysis to detect glottal activity in speech, which is a critical component in emotion expression [15]. By analyzing the Hurst exponent variations, this analysis accurately separated voiced and unvoiced speech segments, which can improve emotion recognition.

The Volterra filter is based on the Volterra series, which can model nonlinear systems with memory. This is beneficial for capturing the intricate patterns in emotional speech. Agrawal and Patil combined nonlinear residual phase features, obtained with a Volterra-Wiener filter with Mel Frequency Cepstral Coefficients (MFCC) for speaker verification tasks [16]. Their study demonstrated improved performance over MFCC alone, indicating the effectiveness of Volterra-based features in capturing speaker-specific information. Daneshfar et al. implemented a second-order Volterra series convolution as a custom TensorFlow layer to improve the precision of speech recognition models and used particularly for speech commands recognition tasks [17]. This approach revealed the potential of

integrating Volterra-based methods into deep learning frameworks for improved speech processing. Roheda and Krim proposed Volterra Neural Networks (VNNs). They used Volterra filter-inspired architectures to introduce controlled nonlinearities in neural networks [18]. Despite their work focused on action recognition, the underlying principles of VNNs could be adapted for SER tasks.

No prior work has been reported to combine Volterra-based nonlinear modeling with multifractal analysis in a single model for speech emotion. This proposed combination of feature set and filtering has the potential to capture both memory-driven nonlinear interactions i.e., through Volterra modelling and scale-invariant complexity through multifractal analysis. Thus, this combination provides a more robust feature representation for emotional speech classification.

Toward this goal, a hybrid SER framework is proposed. It fuses MFCCs, multifractal features, and prosodic cues [19], enhanced using a Sparse Adaptive Volterra Filter, which captures higher-order nonlinear interactions within speech. A CNN-BiLSTM model is used for classification. It exploits both spatial and temporal discriminative capabilities. The proposed system evaluated on a Malayalam Emotional Speech Dataset consisting of 3120 utterances and further validated on the RAVDESS dataset [20].

2 Methodology

The proposed Speech Emotion Recognition (SER) system adopts a hybrid approach combining linear and nonlinear features and a deep learning-based classifier, CNN-BiLSTM. The architecture explores various combinations of four key stages: sparse adaptive Volterra filtering, feature extraction, feature fusion strategies, and classification using models such as CNN, BiLSTM, and their hybrid CNN-BiLSTM configuration.

2.1 Feature extraction

To comprehensively capture the characteristics of emotional speech, three categories of features were extracted:

- MFCCs**:** Spectral features modeling perceptual aspects of human auditory processing, effective for encoding linear acoustic information.
- Multifractal Features: Derived using CWT-Based Structure-Function Multifractal Analysis, capturing scale-invariant, nonlinear irregularities in speech signals associated with emotion.
- Prosodic Features: Including pitch, energy, and formant frequencies, carrying essential cues about emotion such as arousal and valence.

2.2 Sparse Adaptive Volterra Filter (SAVF)

The extracted features are enhanced using a Sparse Adaptive Volterra Filter, a nonlinear system that models higher-order interactions in the speech signal

[21−25]. The sparse variant retains only the most informative Volterra kernel terms, improving computational efficiency while capturing subtle nonlinear dependencies among features [26].

Volterra Filter

The Volterra filter extends linear filters to model nonlinear, time-invariant systems, making it suitable for speech signals with formant interactions and co-articulation effects. Its output is given by:

$$y[n] = h_0 + \sum_{k=0}^{M} h_1[k]x[n-k] + \sum_{k=0}^{M}\sum_{l=0}^{M} h_2[k,l]\,x[n-k]\,x[n-l]\,. \quad (1)$$

Typically, the Volterra series is truncated at the second order to balance nonlinearity modeling with computational efficiency.

Sparse Adaptive Volterra Filter (SAVF)

To reduce complexity, Sparse Adaptive Volterra Filters (SAVF) activate only a subset of second-order interactions using a time-varying sparse mask [26]:

$$y[n] = h_0 + \sum_{k=0}^{M} h_1[k]x[n-k] + \sum_{k=0}^{M}\sum_{l=0}^{M} m[n;k,l]h_2[k,l]x[n-k]x[n-l]\,. \quad (2)$$

$m[n;k,l] \in \{0,1\}$ selects the active interactions, with $S \ll (M+1)(M+2)/2$ nonzero terms.

Advantages of SAVF

– Efficiency: Reduces computational load by retaining only significant nonlinear terms.

– Accurate Modeling: Focuses on informative interactions, minimizing overfitting [27].

– Adaptivity: Effectively tracks nonstationary speech characteristics like articulation dynamics and speaker variations [28 − 31].

2.3 Multifractal analysis

The CWT-based structure-function multifractal analysis method is used to extract multifractal features. The corresponding algorithm is detailed below [32].

Algorithm: CWT-Based Structure-Function Multifractal Analysis

Input: Signal $x(t)$, complex wavelet ψ, scale set A, moment orders Q.

Output: Multifractal descriptor vector *mfd*.

1. Preprocess the signal:
   – Normalize and trim silence.
2. Compute the CWT: Obtain coefficients $W(a,b)$.
3. Stabilize modulus: $|W(a,b)| = \max(|W(a,b)|, 10^{-12})$.

4. Compute structure functions: For each $q \in Q$ and scale $a \in A$, compute

$$S_q(a) = \begin{cases} \exp\left(\langle \ln |W(a,b)| \rangle_b\right), & q = 0, \\ [6\,\text{pt}]\langle |W(a,b)|^q \rangle_b^{1/q}, & q \neq 0. \end{cases}$$

5. Estimate scaling exponent: Fit $\log S_q(a)$ vs. $\log a$ to get slope $\tau(q)$

6. Compute singularity strength: $h(q) = \dfrac{\mathrm{d}\,\tau(q)}{\mathrm{d}\,q}$.

7. Legendre transform: $\alpha(q) = h(q), \quad f(\alpha(q)) = q\alpha(q) - \tau(q)$.

8. Package descriptor: $\text{mfd} = \{\tau(q), h(q), \Delta h, f(\alpha_{\max})\}$.

The proposed multifractal feature extraction starts with pre-processing the audio. Each signal is normalized to consistent amplitude. The silent parts at the beginning and end are trimmed. This step ensures that the analysis concentrates only on the active speech.

Then, Continuous Wavelet Transform (CWT) is applied. This split signal down into time and frequency components. To keep the calculations stable, a small threshold is applied to the wavelet coefficients particularly for very small values to avoid problems during processing.

Next, structure functions are computed which examines the signal behavior across different scales. These functions are calculated for a range of values called q-values, each showing a different type of variations in the signal. Higher q-values bring out large variations. Lower q-values emphasize small variations. By studying these variations with scale, scaling exponents are computed, which describes the self-similarity across the signal.

The algorithm then calculates singularity strengths, which is a way to measure smoothness or irregularity of different parts of the signal. A wider range of these values indicates a more complex structure, which is often seen in emotionally expressive speech. The multifractal spectrum shows how these variations are distributed across the signal.

As the final step, the algorithm combines all these features into a set of descriptive features such as the width of the multifractal spectrum, full range of scaling and singularity, and the peak of the spectrum. This combined feature set is then used as input to the CNN-BiLSTM to classify emotions.

2.4 Prosodic Features

Pitch contours are extracted using the PYIN algorithm [23]. Formant frequencies are estimated using Linear Predictive Coding (LPC) to extract the resonant properties of the vocal tract [24]. The mean value was calculated for both pitch and each formant trajectory to represent the central tendency and the

standard deviation was computed to measure variability. These values give a compact summary of prosodic descriptors.

The stepwise fusion strategy helps identify the optimal combination of features for emotion recognition. The combined feature set is shown in Fig. 1.

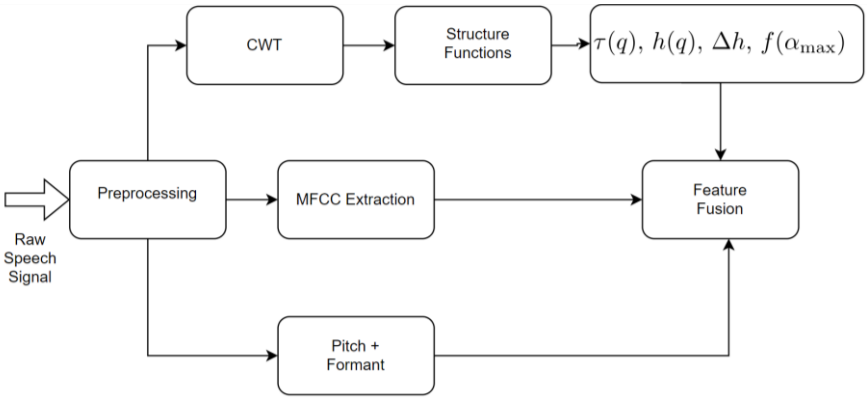


Fig. 1 – *Combined feature set used in the proposed SER method.*

3 Experimental Setup

3.1 Dataset Description

A new emotional speech dataset was recorded for this study. The dataset was built from 12 sentences, consisting of 2 common sentences and 12 phonetically balanced ones, resulting in a total of 42 sentences. Each sentence was recorded in all four emotions for consistency and repeated three times. The dataset includes recordings from 22 speakers, with an equal number of male and female participants (11 each). Due to recording constraints, two speakers recorded only 10 sentences per emotion.

The recordings are captured in a professional studio environment using a Neumann TLM 102 microphone, with a sampling rate of 48 kHz and a bit depth of 24 bits. The software used for recording is Pro One.

The total number of samples was calculated as follows:

– 20 standard speakers recorded 12 sentences per emotion, 3 repetitions per sentence, resulting in 2,800 samples.

– 2 partial speakers recorded 10 sentences per emotion, 3 repetitions per sentence, yielding 240 samples.

The dataset included 3120 utterances across four emotions (Happy, Sad, Angry, and Neutral). Dataset expanded to 12,480 samples using augmentation

techniques. The augmented version was created using techniques such as Gaussian noise addition (noise amplitude: 0.001-0.015), pitch shifting (± 2 semitones), time stretching (0.8-1.25x), and time shifting (-50% to +50%).

Experiments were carried out on three datasets.

Same-sentence dataset: This dataset consists of 480 utterances of two common sentences in four emotions by 20 speakers with three repetitions each. After augmentation, it contains 1920 samples.

Emotional Speech Dataset: This dataset consists of two common and 10 phonetically balanced sentences. It expanded from 3120 to 12,480 samples after augmentation.

RAVDESS Dataset: The publicly available RAVDESS dataset was used for evaluation. This dataset consists of 1,440 utterances across eight emotions. There are 192 samples per emotion, except the neutral class, which has only 96 samples. A subset of four emotions (Neutral, Sad, Happy, and Angry) was selected to match the emotions in the dataset recorded for this study. Both the original and fourfold augmented versions of the dataset were used for the validation.

All these datasets include different speakers with varied emotional expressions and recorded without background noise.

3.2 Classification

The Baseline CNN architecture consists of three 1D convolutional blocks with 64, 128, and 256 filters, each followed by ReLU activation, batch normalization, and maxpooling to progressively extract hierarchical features. The convolution layers are followed by a global average pooling layer, which is employed to aggregate the temporal dimensions after the convolutional layers to significantly reduce the feature map size. Then a fully connected dense layer with 128 units, and a dropout of 0.5 is applied to prevent overfitting. Then a final softmax output layer and the model is optimized using the Adam optimizer, learning rate is set as 0.001 and trained with a batch size of 16.

The CNN–BiLSTM Hybrid model integrates convolutional and recurrent layers. It captures both local and long-term temporal dependencies. There are two convolutional layers having 64 and 128 filters, respectively, to extract local features. A two stacked Bidirectional LSTM layers with 128 and 64 units follows enabling the model to learn contextual temporal relationships in both forward and backward directions. A dense layer with 64 units with a dropout rate of 0.3 applied to mitigate overfitting is used before the softmax output layer. The Adam optimizer with a learning rate of 0.001 and a batch size of 16 is used for training.

Multiple combinations of features evaluated to assess the impact of each feature type on classification accuracy.

To evaluate the CWT-Based Structure-Function Multifractal Analysis method, comparisons were conducted with two widely used multifractal techniques, WTMM and MF-DFA [8].

4 Results and Discussion

The models are evaluated using accuracy, precision, and recall, defined as:

– Accuracy: Measures the overall correctness of predictions

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3}$$

– Precision: Proportion of correctly predicted positive observations to the total predicted positives

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{4}$$

– Recall (Sensitivity): Proportion of correctly predicted positive observations to all actual positives

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{5}$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives.

4.1 Results

Table 1 compares the accuracy of WTMM, MF-DFA, and the CWT-Based Structure-Function Multifractal Analysis approach for emotion recognition.

The comparison of accuracy across various methods reveals notable performance differences, as shown in the **Table 1**. Among the individual methods, proposed method of multifractal analysis achieves an accuracy of 55.2%, WTMM at 47.25% and MF-DFA at 38.71%. Specifically, when MFCC features are added to each method, the accuracy improves, with MF-DFA+MFCC reaching 80.02% and WTMM+MFCC achieving 90.75%. However, the most significant improvement is observed with the Proposed Method, which achieves an accuracy of 55.2%, and when combined with MFCC features, it achieves the highest accuracy of 91.03%. This shows that the proposed method, especially the combination of features with combined with MFCC features, gave higher than other methods.

– The classification accuracy of approximately 88.40% achieved with CNN model using MFCC features.

– The CNN-BiLSTM model further improved classification accuracy to 89.10, effectively capturing both spatial and temporal patterns in speech signals.

Table 1
Performance Metrics for Different Multifractal Feature Extraction Methods for Emotion Recognition.

Methods	Accuracy	Precision	Recall	F1-score
MF-DFA	38.71	0.38	0.38	0.38
WTMM	47.25	0.47	0.47	0.47
CWT-Based Structure-Function Multifractal Analysis	55.2	0.55	0.55	0.55
MFCC	89.10	0.89	0.89	0.89
MF-DFA+MFCC	80.02	0.80	0.79	0.80
WTMM+MFCC	90.75	0.93	0.93	0.94
CWT-Based Structure-Function Multifractal Analysis+MFCC	91.03	0.91	0.90	0.91

Fig. 2 illustrates the classification accuracy for different combinations of features. "MF" refers to Multifractal features, "PF" to Pitch and Formant features, and "V" to Sparse Adaptive Volterra Filter-based features. Accuracy increases as each feature type is added to the MFCC features. The highest accuracy is reached all feature types are combined.

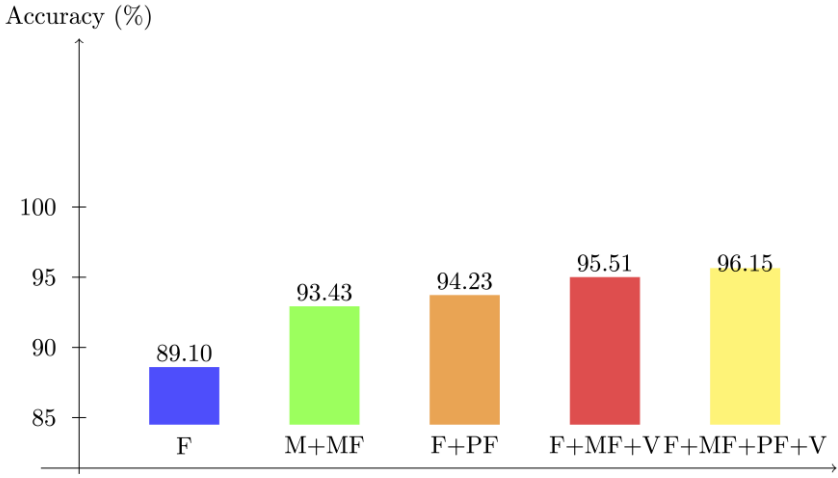


Fig. 2 – *Accuracy Comparison, F = MFCC, MF = Multifractal, PF = Pitch & Formants, V = Sparse Adaptive Volterra Filter.*

4.2 Experiments on original and augmented datasets

Initially, MFCC features are extracted. MFCC feature extraction was used to train the emotion recognition models. The CNN model trained on the original

dataset achieved an accuracy of 89.10%, while training on the augmented dataset achieved an accuracy of approximately 91.03%, demonstrating the effectiveness of data augmentation.

Then MFCC is combined with multifractal features, and classification is performed. CNN and CNN-BiLSTM classifiers were used in the experimental setup. The classification accuracy achieved on the original dataset was 89.74% with CNN and 93.43% with CNN-BiLSTM After applying data augmentation to the original dataset, the accuracy improved to 97.44% with CNN and 97.28% with CNN-BiLSTM.

As a next step, a Sparse Adaptive Volterra Filter was applied to the audio signals to enhance nonlinear feature representation of the signals. Then, MFCC and multifractal features were extracted from these filtered signals, and classification was repeated with the same classifiers. This resulted in improved accuracy of 93.75% with CNN and 95.51% with CNN-BiLSTM on the original dataset.

As the final step, pitch and formant features are computed and appended to the existing MFCC and multifractal features. Also, Sparse Adaptive Volterra Filter is used. Classification using this extended feature set yielded the highest accuracy of 93.43% with CNN and 96.15% with CNN-BiLSTM. The results in **Table 2** demonstrate the effectiveness of integrating MFCC, prosodic, and nonlinear acoustic features for robust speech classification. Moreover, the performance remained promising on the extended dataset as well, as shown in **Table 2**. Since two sentences were common across all emotions, the results for this same-sentence subset using different feature combinations and data augmentation are presented in **Table 1**.

Table 2
Accuracy (%) *on Full In-House Dataset.*

Features	CNN	CNN + BiLSTM
MFCC	88.40	89.10
MFCC + MF	89.74	93.43
MFCC + Pitch & Formants	92.47	94.23
MFCC + MF + V	93.75	95.51
MFCC + MF + Pitch & Formants + V	93.43	96.15
MFCC (Augmented)	92.88	95.63
MFCC + MF (Augmented)	97.44	97.28
MFCC + MF +V (Augmented)	97.72	98.08

Incorporating the Sparse Adaptive Volterra filter along with MFCC+MF and pitch and formants leads to an improvement in accuracy.

To further analyze performance, the confusion matrices shown in Figs. 3 and 4 illustrate the classification results for each class.

Fig. 3 shows the confusion matrix for MFCC+MF on the original dataset. Fig. 4 shows the confusion matrix for MFCC+MF computed on the augmented same-sentence dataset and confusion matrix for MFCC+MF+V computed on the augmented same-sentence dataset.

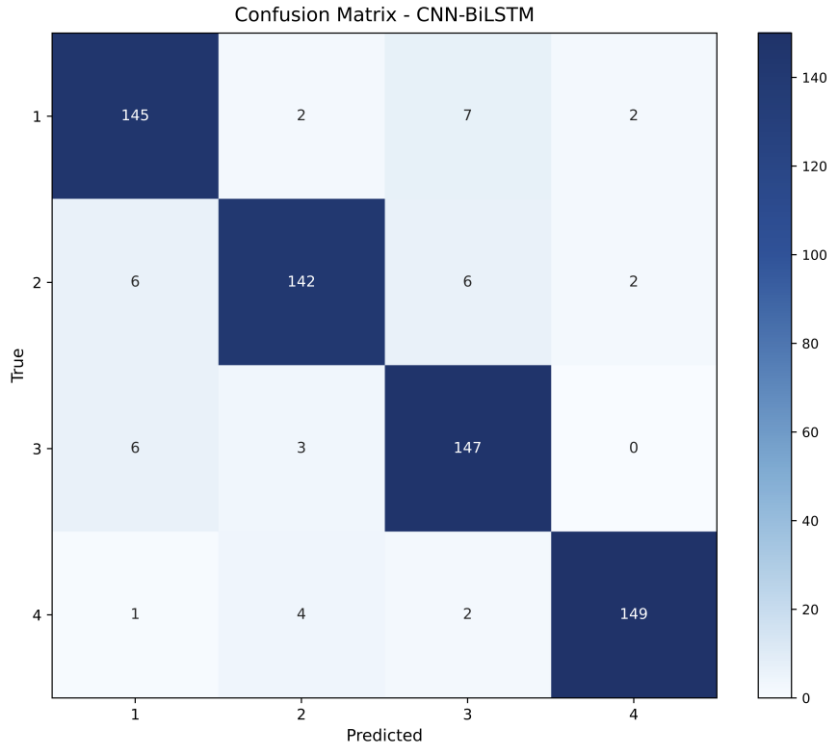
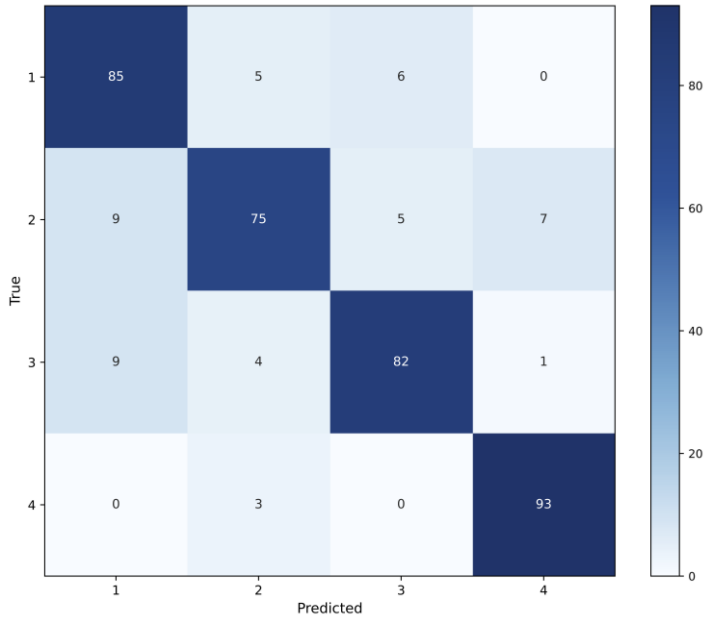
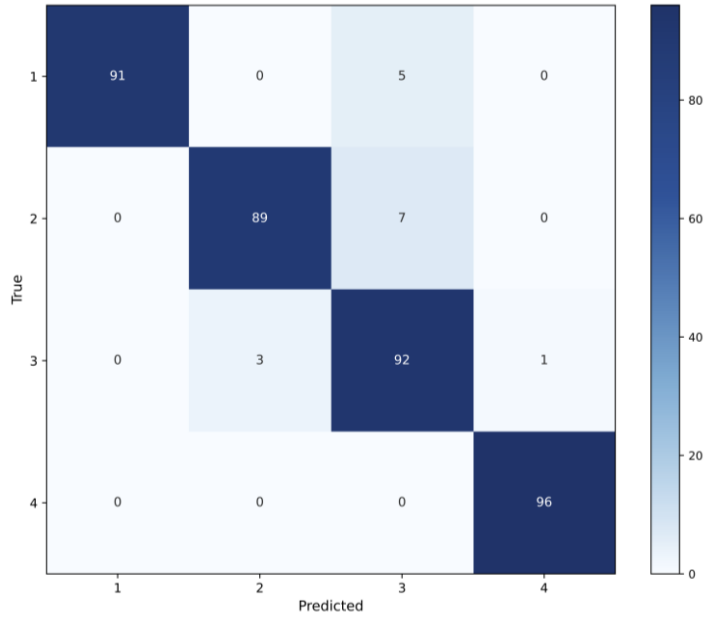


Fig. 3 – *MFCC + MF on original dataset.*

In the same-sentence dataset, the class labels correspond to the following emotions: Class 1 – Neutral, Class 2 – Happy, Class 3 – S ad, and Class 4 – Angry. As shown in **Table 3**, Angry emotion was recognised most accurately in all experiments, followed by Neutral, Happy, and Sad. This suggests that high-arousal emotions like anger are more distinctly captured by the proposed features, whereas low-arousal emotions such as sadness are comparatively harder to classify accurately.



(a)



(b)

Fig. 4 – (a) *Confusion matrix for MFCC+MF on the augmented same sentence dataset;* (b) *Confusion matrix for MFCC+MF+V on the augmented same sentence dataset.*

Table 3
Accuracy (%) *on Same Sentence Dataset.*

Features	CNN	CNN-BiLSTM
MFCC + MF	70.83	71.88
MFCC + MF + V	80.21	85.42
MFCC + MF (Augmented)	86.2	93.75
MFCC + MF + V (Augmented)	95.31	96.35

Table 4
Classification Report for CNN-BiLSTM (MFCC + MF, Same Sentence Dataset).

Class	Precision	Recall	F1-score	Support
1	0.89	0.93	0.91	96
2	0.97	0.91	0.94	96
3	0.92	0.95	0.93	96
4	0.98	0.97	0.97	96
Accuracy			0.94	384
Macro Avg	0.94	0.94	0.94	384
Weighted Avg	0.94	0.94	0.94	384

As shown in **Table 2**, **3** and **4**, and illustrated in Fig. 5, the proposed feature combinations yield consistent improvements in accuracy across different datasets.

The RAVDESS dataset was also used to test the proposed method. The same features were computed from the RAVDESS audio files, and the same classification pipeline was used. Since the RAVDESS dataset contains eight emotions, accuracy was first calculated for 8-class classification and then for the four emotions used in the recorded dataset. The models again achieved high accuracy as shown in **Table 5** and Fig. 5. This suggests that the method also works well on different corpora.

Table 5
Accuracy (%) *on RAVDESS Dataset.*

Features	CNN	CNN + BiLSTM
MFCC + MF (8 Emotions)	86.16	73.51
MFCC + MF (4 Emotions)	89.63	77.78
MFCC + MF (Augmented 8 Emotions)	96.09	94.79
MFCC + MF (Augmented 4 Emotions)	96.47	95.91

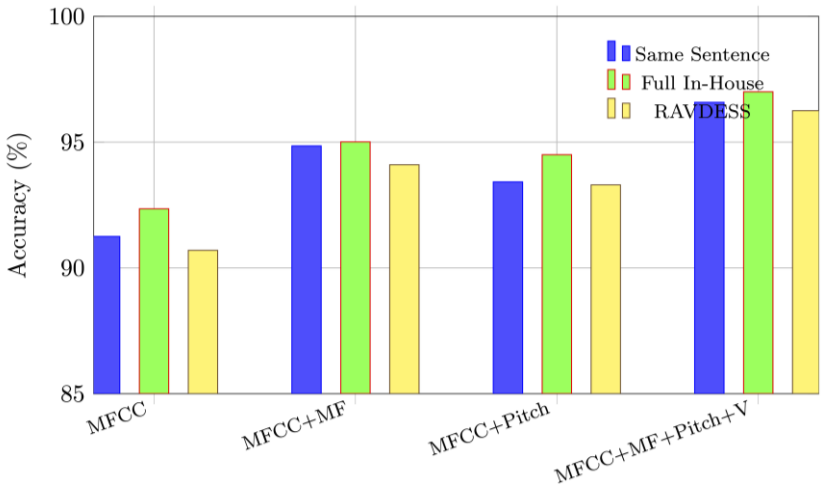


Fig. 5 – *Accuracy comparison across feature combinations and datasets.*

The proposed feature extraction pipeline includes preprocessing, MFCC computation, multifractal analysis. For a signal of length T, preprocessing and normalization take O (T) time. MFCC extraction takes approximately O (T logT). The multifractal stage computes the structure functions for Q moment orders across S scales, which requires O(QST) time. This is the most expensive step. Feature fusion adds only a small cost, proportional to the number of frames. On average, this corresponds to 0.498 s per sample for feature extraction and 0.313 s per sample for classification, yielding an overall processing time of 0.811 s per sample and an overall average processing time of 0.3536 s per file for MFCC.

5 Conclusion

The proposed framework combines MFCC, multifractal features, prosodic cues and sparse adaptive Volterra filtering with a CNN and a bidirectional LSTM model. It achieved high accuracy on both the Malayalam in-house dataset (expanded from 3120 to 12480 samples by augmentation) and the RAVDESS benchmark. Adding multifractal information to MFCC improved emotion classification, raising accuracy from 89.10% to 93.43%, and the Volterra filter increased it further to 95.51%. These results support the advantage of combining spectral, nonlinear, and prosodic characteristics of speech emotion recognition.

In future work, feature selection methods such as PCA and LASSO will be applied to identify the most informative features. On the model side, separate network branches will be developed for each feature type, with an attention mechanism to combine their outputs**.** The proposed method will be tested on more emotional speech datasets, and its feasibility for real-time deployment will be evaluated.

6 References

[1] M. El Ayadi, M. S. Kamel, F. Karray: Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases, Pattern Recognition, Vol. 44, No. 3, March 2011, pp. 572 – 587.

[2] D. Ververidis, C. Kotropoulos: Emotional Speech Recognition: Resources, Features, and Methods, Speech Communication, Vol. 48, No. 9, September 2006, pp. 1162 – 1181.

[3] S. Davis, P. Mermelstein: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28, No. 4, August 1980, pp. 357 – 366.

[4] H. Kantz, T. Schreiber: Nonlinear Time Series Analysis, Cambridge University Press, Cambridge, 2004.

[5] E. A. F. Ihlen: Introduction to Multifractal Detrended Fluctuation Analysis in MATLAB, Frontiers in Physiology, Vol. 3, June 2012, p. 141.

[6] J. F. Muzy, E. Bacry, A. Arneodo: Wavelets and Multifractal Formalism for Singular Signals: Application to Turbulence Data, Physical Review Letters, Vol. 67, No. 25, December 1991, pp. 3515 – 3518.

[7] S. G. Koolagudi, K. S. Rao: Emotion Recognition from Speech: A Review, International Journal of Speech Technology, Vol. 15, No. 2, January 2012, pp. 99 – 117.

[8] S. Bhaduri, A. Bhaduri, R. Sarkar: Language Independent Speech Emotion and Non-Invasive Early Detection of Neurocognitive Disorder, arXiv:2106.01684v1 [cs.SD], June 2021, pp. 1 – 10.

[9] S. Paul, A. Mazumder, P. Ghosh, D. N. Tibarewala, G. Vimalarani: EEG Based Emotion Recognition System Using MFDFA as Feature Extractor, Proceedings of the International Conference on Robotics, Automation, Control and Embedded Systems (RACE), Chennai, India, February 2015, pp. 1 – 5.

[10] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, IEEE Signal Processing Magazine, Vol. 29, No. 6, November 2012, pp. 82 – 97.

[11] S. Zhang, S. Zhang, T. Huang, W. Gao: Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching, IEEE Transactions on Multimedia, Vol. 20, No. 6, June 2018, pp. 1576 – 1590.

[12] U. Sarkara at al: Language Independent Emotion Quantification using Non-linear Modelling of Speech, arXiv:2102.06003, 2021.

[13] X. Mao, L. Chen: Speech Emotion Recognition Based on Parametric Filter and Fractal Dimensions, IEICE Transactions on Information and Systems, Vol. E93.D, No. 8, August 2010, pp. 2324 – 2326.

[14] G. Tamulevičius, G. Korvel, A. B. Yayak, P. Treigys, J. Bernatavičienė, B. Kostek: A study of cross-linguistic speech emotion recognition based on 2D feature spaces, Electronics, Vol. 9, No. 10, October 2020, p. 1725.

[15] G. J. Lal, E. A. Gopalakrishnan, D. Govind: Glottal Activity Detection from the Speech Signal Using Multifractal Analysis, Circuits, Systems, and Signal Processing, Vol. 39, No. 4, pp. 2118 – 2150, Apr. 2020.

[16] P. Agrawal, H. A. Patil: Fusion of a Novel Volterra–Wiener Filter Based Nonlinear Residual Phase and MFCC for Speaker Verification, Proceedings of the International Conference on Speech and Computer (SPECOM), Hatfield, UK, September 2017, pp. 389 – 397.

[17] F. Daneshfar, S. J. Kabudian: Speech Emotion Recognition System by Quaternion Nonlinear Echo State Network, arXiv:2111.07234v1 [cs.SD], November 2021, pp. 1 – 44.

[18] S. Roheda, H. Krim, B. Jiang: Volterra Neural Networks (VNNs), Journal of Machine Learning Research, Vol. 25, No. 182, 2024, pp. 1 – 29.

[19] F. Eyben, M. Wöllmer, B. Schuller: Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor, Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, October 2010, pp. 1459 – 1462.

[20] S. R. Livingstone, F. A. Russo: 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, Vol. 13, No. 5, May 2018, p. e0196391.

[21] A. Rai, A. Kumar Kohli: Volterra Filtering Scheme Using Generalized Variable Step-Size NLMS Algorithm for Nonlinear Acoustic Echo Cancellation, Acta Acustica United with Acustica, Vol. 101, No. 4, 2015, pp. 821 – 828.

[22] L. A. Azpicueta-Ruiz, M. Zeller, A. R. Figueiras-Vidal, J. Arenas-Garcia, W. Kellermann: Adaptive Combination of Volterra Kernels and Its Application to Nonlinear Acoustic Echo Cancellation, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 1, January 2011, pp. 97 – 110.

[23] Y. Mu, P. Ji, W. Ji, M. Wu, J. Yang: Modeling and Compensation for the Distortion of Parametric Loudspeakers Using a One-Dimensional Volterra Filter, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 22, No. 12, December 2014, pp. 2169 – 2181.

[24] C. Shi, Y. Kajikawa: Volterra Model of the Parametric Array Loudspeaker Operating at Ultrasonic Frequencies, The Journal of the Acoustical Society of America, Vol. 140, No. 5, November 2016, pp. 3643 – 3650.

[25] T. Roy, P. Ahmmed, M. S. Rahman, A. Ash-Saki, A. B. M. Harun-Ur-Rashid: A Novel Adaptive Volterra Filter to Compensate for Speaker Non-Linearity, Proceedings of the 8th International Conference on Electrical and Computer Engineering, Dhaka, Bangladesh, December 2014, pp. 258 – 261.

[26] M. Schetzen: The Volterra and Wiener Theories of Nonlinear Systems, John Wiley & Sons, New York, 1980.

[27] M. Zeller, L. A. Azpicueta-Ruiz, J. Arenas-García, W. Kellermann: Adaptive Volterra Filters with Evolutionary Quadratic Kernels Using a Combination Scheme for Memory Control, IEEE Transactions on Signal Processing, Vol. 59, No. 4, April 2011, pp. 1449 – 1464.

[28] J. Arenas-García, L. A. Azpicueta-Ruiz, M. T. M. Silva, V. H. Nascimento, A. H. Sayed: Combinations of Adaptive Filters: Performance and Convergence Properties, IEEE Signal Processing Magazine, Vol. 33, No. 1, January 2016, pp. 120 – 140.

[29] S. Xu, Y. Li, T. Huang, R. H. M. Chan: A Sparse Multiwavelet-Based Generalized Laguerre–Volterra Model for Identifying Time-Varying Neural Dynamics from Spiking Activities, Entropy, Vol. 19, No. 8, August 2017, p. 425.

[30] M. Mauch, S. Dixon: PYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy, May 2014, pp. 659 – 663.

[31] L. R. Rabiner, R. W. Schafer: Digital Processing of Speech Signals, Prentice-Hall, Inc., Englewood Cliffs, 1978.

[32] F. K. Mohamed, P. P. Abdul Haleem: Enhancing Malayalam Speech Recognition Through Multifractal Analysis, International Journal of Speech Technology, Vol. 29, No. 1, March 2026, p. 36.