Communications in Mathematics and Applications
Vol. 14, No. 3, pp. 1189–1198, 2023
ISSN 0975-8607 (online); 0976-5905 (print)
Published by RGN Publications
DOI: 10.26713/cma.v14i3.2389

http://www.rgnpublications.com

Special Issue
Recent Trends in Mathematics and Applications
Proceedings of the International Conference of Gwalior Academy of Mathematical Sciences 2022
***Editors*: Vinod P. Saxena and Leena Sharma**

Research Article

Performance Improvement in Speech Based Emotion Recognition With DWT and ANOVA

Ashwini S. Shinde*[1,2] and Vaishali V. Patil[3]

[1] *E&TC Department, AISSM's Institute of Information Technology* (*Savitribai Phule Pune University*), *Pune, India*
[2] *E&TC Department, Pimpri Chichwad College of Engineering* (*Savitribai Phule Pune University*), *Pune, India*
[3] *E&TC Department, International Institute of Information Technology* (*Savitribai Phule Pune University*), *Pune, India*
***Corresponding author:** ashwinik09@gmail.com

Received: February 2, 2023 **Accepted:** June 4, 2023

Abstract. With technological advancements, machine needs to understand human speech, i.e., *Human-Computer Interaction* (HCI) has become vital. For natural interaction, emotion detection in speech is a must. Time domain features can identify a few emotions, whereas some are determined by inherently using frequency domain features. With wavelet-based features majority of emotion-discriminating features can be identified. A common observation is that happy emotion is seen to be majorly misclassified as an angry emotion. Reduction in this misclassification is achieved with the proposed feature vector. Spectral features and *Discrete Wavelet Transform* (DWT) features form the proposed feature vector. Feature selection is made using statistical test analysis of variance ANOVA. The model is verified using SVM and MLP classifiers. In this work, a speech emotion recognition system is evaluated using a German audio database (EMODB). It is seen to be able to recognize happy and angry emotions with better accuracy as compared to state-of-the-art algorithms. For four emotion classes: happy, angry, neutral and sad proposed model performance with DWT features has improved by 3% compared to baseline features in the case of both the classifiers, viz. SVM and MLP.

Keywords. Speech emotion recognition, DWT, ANOVA, SVM, MLP

Mathematics Subject Classification (2020). 62P30, 62J10, 65T60, 68T07

1. Introduction

Every human can comfortably express his state of mind by communicating full of emotions. *Speech Emotion Recognition* (SER) is the act of attempting to acknowledge human emotion and affective states from speech. An SER model has been developed to assist people to identify the emotions of people. Choice of a dataset, discriminating feature extraction algorithms, reducing feature space and optimizing performance feature selection, and finally selection of classifier is challenging to work upon (Shinde and Patil [9]). El Ayadi *et al*. [4] and Swain [12] present an extensive and detailed literature study with methodology, challenges, and research gaps. At present, many researchers are focusing on improving performance with natural speech emotion recognition databases. Selection of different features, a variety of feature selection methods, and the choice of a suitable classifier is a challenging tasks in emotion detection based on speech. Many researchers have explored spectral characteristics describing spectral features such as *Mel Frequency Cepstral Coefficients* (MFCC), Bark scale-based *Gammatone Frequency Cepstral Coefficients* (GFCC), bark scale and *Equivalent Rectangular Bandwidth* (ERB) scale based features (Sugan *et al*. [11]). Various feature combinations such as spectral, prosodic, and vocal tract based features always proved better in comparison with baseline features (Deb and Dandapat [3], and Shinde *et al*. [10]). To localize the signal in both the time domain and frequency domain, discrete wavelet transform-based features are also explored.

Due to better performance in speech emotion recognition Deb and Dandapat [3] used Daubachies (db4) as a basis function from the discrete wavelet transform family. They derived new amplitude features based on each sub band of wavelet. They used several databases, out of which for standard Berlin EMODB free database recognition average accuracy for speaker-independent experiment with 39 MFCCs achieved is 76.6% whereas with new multi-resolution amplitude-based features increased to 83.8% (Burkhardt *et al*. [2]). Abdel-Hamid [1] extracted various types of features along with four-level decomposition discrete wavelet-derived entropy features. Experimentation is done on the Arabic speech emotion database with four classes: Angry (A), Happy (H), Sad (S), and Neutral (N) for speaker-independent experiments angry emotions are well identified as compared with happy. Saste and Jagdale [7] used feature vectors created by MFCC and statistical features derived from three-level wavelet decomposition. They used special scared emotion along with other basic emotions for security applications. *Discrete Wavelet Transform* (DWT) is used for the classification of audio signals and specific frequency-related noise removal.

The proposed work uses EMODB dataset, *Support Vector Machine* (SVM) and the *Multilayer Perceptron* (MLP) classifiers to train, recognize and classify emotions and achieved an accuracy of 90% with SVM. The different emotions used are Angry (A), Happy (H), Neutral (N), and Sad (S). The highlights of this proposed work are listed as:

(i) Use of statistical features derived from Db4 discrete wavelet decomposition.

(ii) Use of statistical test ANOVA for identification of emotion-discriminating features.

(iii) Performance improvement in speech emotion recognition average accuracy especially in happy emotion comparatively with other emotions out of four classes (A,H,N,S) for EMODB dataset using machine learning classifier SVM as well as MLP classifier.

The organization of the paper is as follows: Section 2 discusses the dataset, pre-processing with extracted features, and selection with classification. Section 3 describes the performance of the proposed work with different classifier along with comparisons with state of art work of the proposed work. Section 4 summarizes the results with the future scope of work.

2. Materials and Methods

This section describes the dataset used for experimentation. Along with preprocessing stages, proposed feature extraction, feature selection, and classifier.

2.1 Dataset

We have used open source acted German language database EMODB. For experimentation, four major emotion classes: Angry (A), Happy (H), Sad (S), and Neutral (N) are considered. EMODB is acted emotion database, i.e., emotions are uttered by people with professional acting experience. Database is summarized in Table 1.

Table 1. Database used for proposed work

Name of database	Number of speakers	Number of utterances	Emotion classes
EMODB [2]	10	339	4 (Anger, Happy, Sad, Neutral)

2.2 Preprocessing

In speech processing, preprocessing stages comprise of

- *Pre-Emphasis Filtering* — To maximize the signal-to-noise ratio, the amplitude of a high-frequency signal is boosted with the help of first order high pass filter. The filter equation is eq. (2.1) with $\alpha = 0.97$

$$H[n] = S[n] - \alpha * S[n-1]. \tag{2.1}$$

- *Framing* — To converts non-stationary signal into, a stationary one input signal $S[n]$ is sampled and divided into 25 ms duration frames with 10 ms overlap.
- *Windowing* — Hamming window is used. Eq. (2.2) represents Hamming window

$$v[n] = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N}\right), \quad \text{where } 0 \le n \le N. \tag{2.2}$$

- A pre-emphasized speech sample $h[n]$ is multiplied with a hamming window $v[n]$ then it is given to the feature extraction stage.

2.3 Feature Extraction

Baseline Feature Vector (f_{baseline}): Next stage is feature vector formation. We used the Python LIBROSA module for removing leading, trailing silence and extraction of spectral features from a given audio file. The baseline feature vector comprises of the mean of 26 MFCC, one ZCR, twelve chroma, Mel spectrogram coefficients. The size of the baseline feature vector is 154.

- *MFCC*: The *Mel Frequency Cepstral Coefficient* (MFCC) feature is proven for speech-based emotion detection. 13 MFCCs and 13 Delta MFCCs are extracted.

- *Chroma*: Chroma-based features describe the pitch-related classes. In emotion detection, pitch profile helps to differentiate emotion classes.
- *ZCR*: The *Zero-Crossing Rate* (ZCR) is used to identify speech and non-speech, i.e., noise in the given signal. For the identification of emotions in a speech this quality is beneficial. Emotion wise average of zcr_mean is obtained as shown in Figure 1.

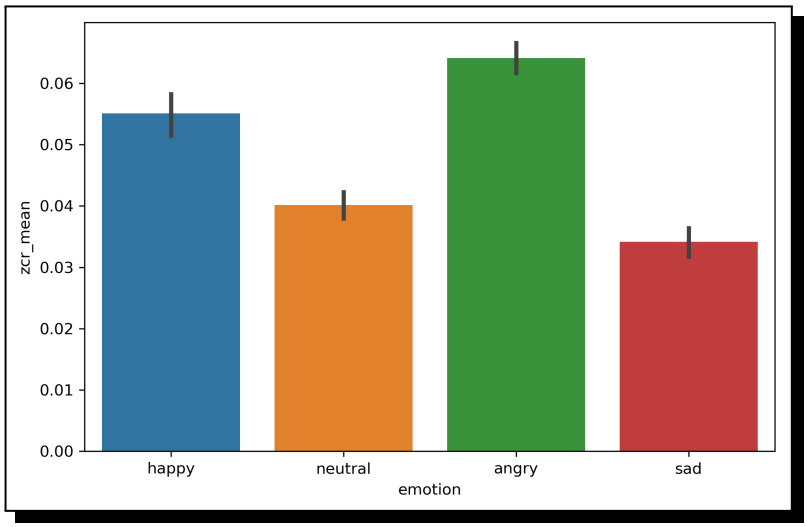

Figure 1. Emotion wise average of zcr_mean

- *Mel spectrogram*: The mel is a frequency scale that mimics the human ear frequency response. A spectrogram is a visual representation of bands of frequencies. The mel frequency scale is defined with

$$\text{mel} = 1125 * \log 10\left(1 + \frac{f}{700}\right). \tag{2.3}$$

Wavelet-based Feature Vector (f_{dwt}): *Discrete Wavelet Transform* (DWT) is used to describe signal in time-frequency domain. The wavelet transform translates a signal's time-amplitude representation to a time-frequency representation of wavelet coefficients. Wavelet transforms are used to derive emotion-describing features [3, 7]. PyWavelets [5] is a Python language package used for sub band decomposition. Using Daubechies4 (db4) wavelet as basis function. Due to decomposition at each level into high pass frequency (cD) and low pass frequency (cA), variation in emotion can be identified. As shown in Figure 2 depicts level-wise decomposition. The original signal can also be reconstructed as represented by eq. (2.4). After the decomposition of audio input into four-level DWT subbands, statistical parameters mean, standard deviation, maximum, and minimum are obtained

$$f_{dwt} = f\{\text{cA4}, \text{cD4}, \text{cD3}, \text{cD2}, \text{cD1}\} \tag{2.4}$$

DWT perform multi sub band resolution. It is represented by

$$T_{m,n} = \int_{-\infty}^{\infty} x(t)\psi_{m,n}(t)dt, \tag{2.5}$$

where $\psi(t)$ is mother wavelet (db4).

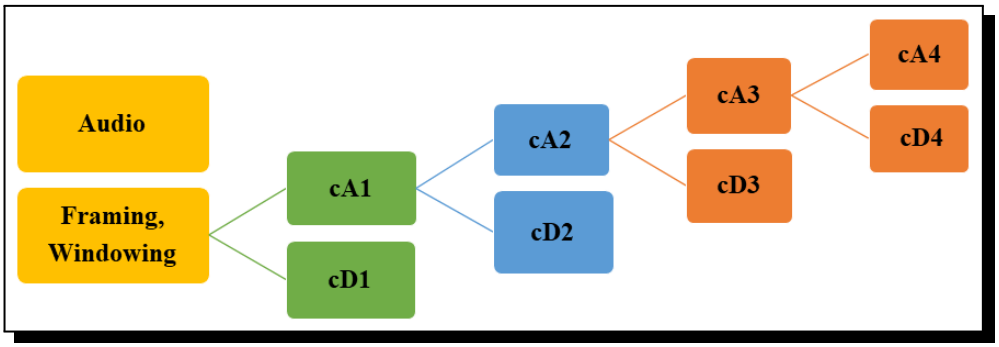

Figure 2. Four-level subband decomposition

One of sample angry audio signal with all four-level subband decomposition is represented in Figure 3. Signal reconstructed by approximation coefficients (cA4) is represented in red color whereas as different subbands reconstructed signal (cD4, cD3, cD2, cD1) are represented using green color. Palo and Mohanty [6] used db4 for three level decomposition of audio derived MFCC and *Linear Prediction Cepstral Coefficients* (LPCC) from wavelet reconstructed signals.

2.4 Feature Selection

Feature selection aims to obtain a subset of discriminative and reduced features from a bigger feature space. Statistical features like Mean, Minimum, Maximum, and Standard Deviation for each subband coefficient, namely cA4, cD4, cD3, cD2, and cD1 of the four-level DWT for all the audio were calculated. The ANOVA (*Analysis of Variance*) statistical test has been used to select the most significant of these statistical features for all the coefficients. Sheikhan *et al*. [8] identified F-score with different combinations of emotion classes. In the proposed work, emotion discriminant features are obtained from 4-level DWT statistical features based on F-score. Combined baseline feature vector and wavelet-based feature vector forms proposed combined feature vector with size 159.

2.5 Classification

In this section, experimentation is carried out with SVM and *Multilayer Perceptron* (MLP) classifiers.

SVM: Support Vector Machine (SVM) creates a hyperplane between various classes. In multi-class classification, either one class versus all other classes or one versus one class scheme is executed. Using kernel functions such as linear, radial basis function, and polynomial, SVM method transforms input data space into the required form of high-dimensional feature space. In this work, the linear kernel function is used to separate and analyze new values based on the training set. SVM with a linear kernel with $C = 1.0$ is used for four classes of emotions. Algorithm has obtained average 90% recognition accuracy with SVM classifier because of the discriminating feature vector.

MLP: Multilayer Perceptron (MLP) classifier is a supervised classification method that uses backpropagation for the training. With the MLP classifier, the recognition accuracy of the proposed work is also verified. Hyperparameters used for MLP are hidden layer = 400, activation function = ReLU and the learning rate is constant with Adam solver.

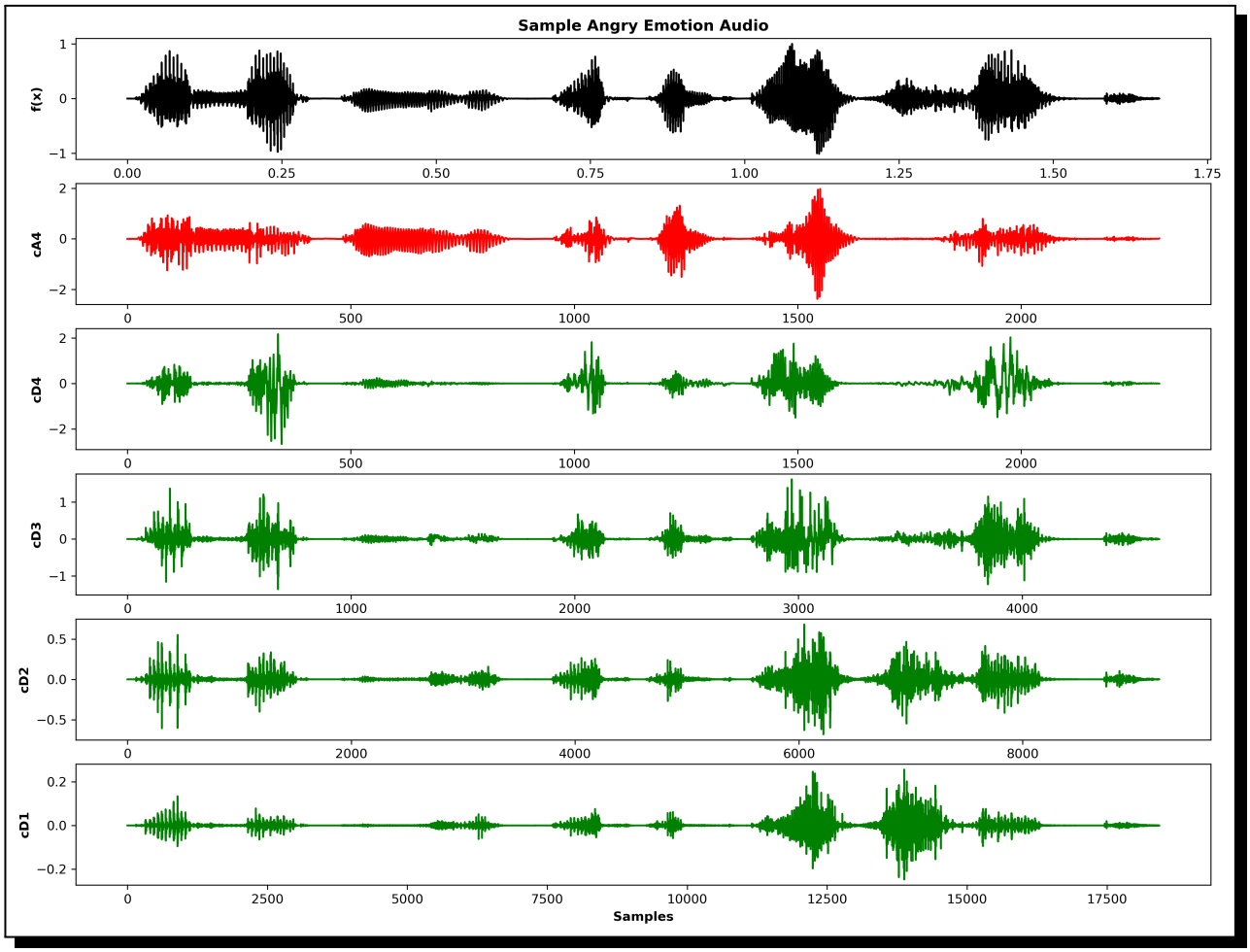

Figure 3. Original signal and its four-levels db4-DWT subbands

Table 2. ANOVA F-score values of discriminative subband features

Feature	cA4-mean	cA4-SD	cA4-min	cA4-max	cA3-SD
F-score	142	180.62	182.4	102.56	102.02

3. Results and Discussion

The baseline features include 154 features of MFCC, ZCR, chroma and mel spectrogram extracted from EMODB. Four-level DWT extracted with EMODB dataset. The statistical features were obtained from DWT coefficients and an ANOVA test was performed on the statistical features and the most significant features were selected. f_{baseline} is the baseline feature vector and f_{dwt} is the wavelet feature vector. A combined feature set $f_{\text{baseline}} + f_{\text{dwt}}$ of 159 features is acquired.

In this section, the confusion matrix of MLP and SVM classifier explains the improvement in average recognition accuracy.

Table 3. Performance (%) for SVM linear with $C = 1.0$ on EMODB

Emotions	Angry	Happy	Neutral	Sad	Average
f_{baseline}	88	71	91	96	**87**
$f_{\text{baseline}} + f_{\text{dwt}}$	88	77	97	100	**90**

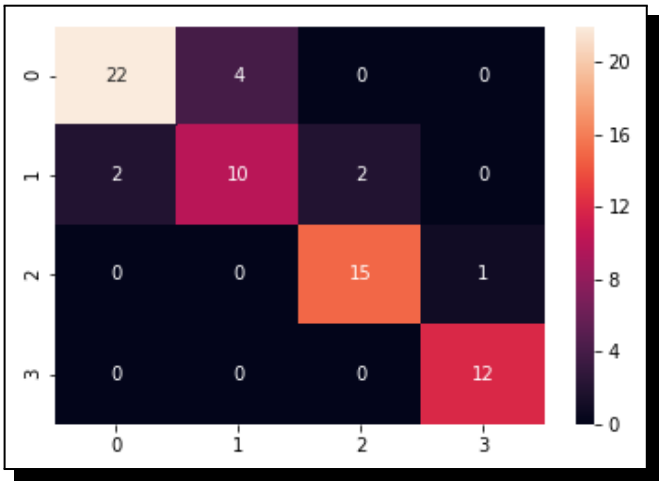

Figure 4. Confusion matrix with SVM linear (EMODB)

Table 3 shows the emotion-wise accuracy of SVM LINEAR classifier. From Table 4 improvement in detection of happy, neutral and sad emotions is observed.

Table 4. Performance (%) with EMODB MLP classifier

Emotions	Angry	Happy	Neutral	Sad	Average
f_{baseline}	85	67	91	92	**84**
$f_{\text{baseline}} + f_{\text{dwt}}$	86	71	94	96	**87**

Table 4 depicts that with MLP classifier; there is an improvement in the recognition rate of emotions happy, neutral, and sad. As represented in Figure 5, misclassification of happy emotion with angry emotion is reduced significantly. Figure 6 represents performance comparison of baseline feature vector and proposed combined feature vector using baseline spectral and DWT-based combined feature vector.

Table 5 depicts state of art comparison with EMODB dataset and proposed work performance.

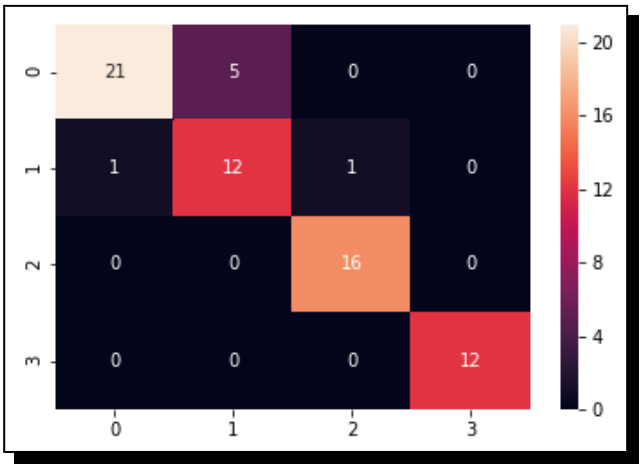

Figure 5. Confusion matrix with MLP classifier (EMODB)

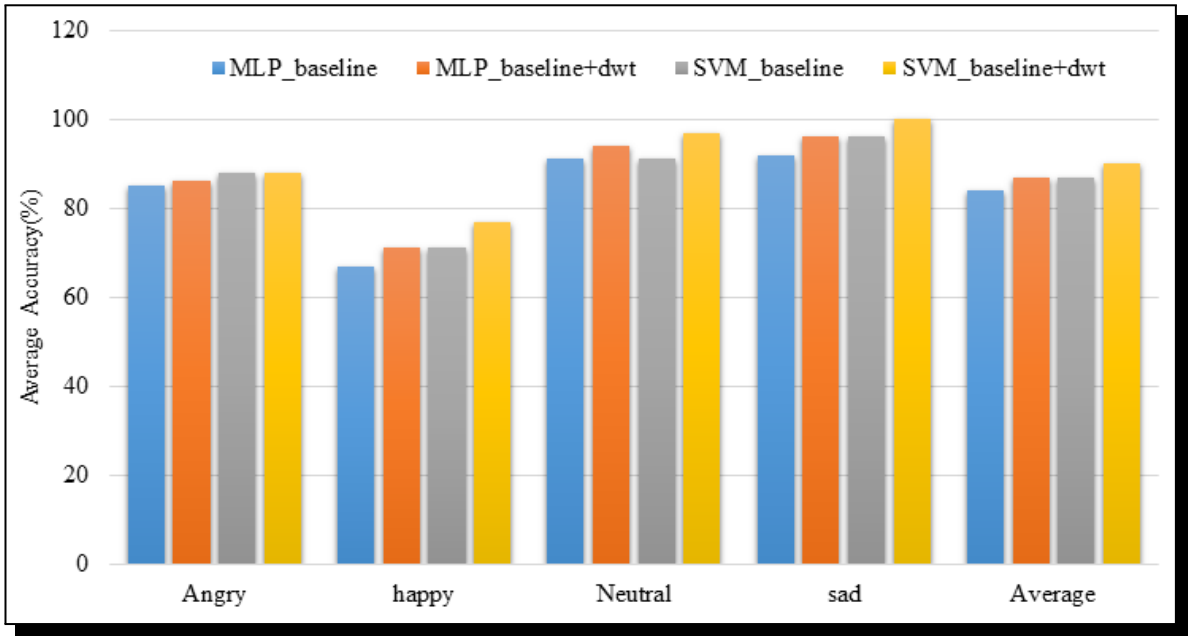

Figure 6. Performance comparison of average recognition accuracy (%) with SVM and MLP on (EMODB) dataset

Table 5. State of art comparison with EMODB dataset

Reference	Methodology	Classifier	(%) Average accuracy
[11]	Spectral features extracted from human factor and mel-scale	SVM	86.96
[3]	Features based on multi-resolution based amplitude and breathiness	SVM	91.60
[7]	Spectral features MFCC	MLP	80.00
Proposed work	Combined feature vector of spectral and DWT features	SVM	**89.70**
		MLP	**87.32**

4. Conclusion

Experiments were conducted on the German EMODB dataset along with SVM and MLP classifiers in speech emotion recognition. With SVM obtained 90% of accuracy performance is better than MLP classifier 87% with baseline feature vector f_{baseline}. With new combined feature vector, i.e., $f_{\text{baseline}} + f_{\text{dwt}}$, average accuracy is improved. As well as with new feature vector, misclassifications in the happy class got reduced significantly. This model can be used in a variety of situations such as voice-based remote helpers or chatbots, an initial diagnostic tool for a psychologist. In the future, work can extend with use of deep learning classifiers, different feature selection methods and different databases.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] L. Abdel-Hamid, Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features, *Speech Communication* **122** (2020), 19 – 30, DOI: 10.1016/j.specom.2020.04.005.

[2] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, A database of German emotional speech, in: *9th European Conference on Speech Communication and Technology*, 2005, pp. 1517 – 1520, DOI: 10.21437/Interspeech.2005-446.

[3] S. Deb and S. Dandapat, Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification, *IEEE Transactions on Cybernetics* **49**(3) (2019), 802 – 815, DOI: 10.1109/TCYB.2017.2787717.

[4] M. El Ayadi, M. S. Kamel and F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition* **44**(3) (2011), 572 – 587, DOI: 10.1016/j.patcog.2010.09.020.

[5] G. R. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt and A. O'Leary, PyWavelets: a Python package for wavelet analysis, *Journal of Open Source Software* **4**(36) (2019), 1237, DOI: 10.21105/joss.01237.

[6] H. K. Palo and M. N. Mohanty, Wavelet based feature combination for recognition of emotions, *Ain Shams Engineering Journal* **9**(4) (2018), 1799 – 1806, DOI: 10.1016/j.asej.2016.11.001.

[7] S. T. Saste and S. M. Jagdale, Emotion recognition from speech using MFCC and DWT for security system, in: *2017 International conference of Electronics, Communication and Aerospace Technology* (ICECA), Coimbatore, India, 2017, pp. 701 – 704, DOI: 10.1109/ICECA.2017.8203631.

[8] M. Sheikhan, M. Bejani and D. Gharavian, Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method, *Neural Computing and Applications* **23** (2013), 215 – 227 (2013), DOI: 10.1007/s00521-012-0814-8.

[9] A. S. Shinde and V. V. Patil, Speech Emotion recognition system: a review, in: *Proceedings of the 4th International Conference on Advances in Science & Technology* (ICAST2021), 2021, 6 pages, DOI: 10.2139/ssrn.3869462.

[10] A. S. Shinde, V. V. Patil, K. R. Khadse, N. Jadhav, S. Joglekar and M. Hatwalne, ML based speech emotion recognition framework for music therapy suggestion system, in: *2022 6th International Conference On Computing, Communication, Control and Automation*, Pune, India, 2022, pp. 1 – 5, DOI: 10.1109/ICCUBEA54992.2022.10011091.

[11] N. Sugan, N. S. S. Srinivas, L. S. Kumar, M. K. Nath and A. Kanhe, Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and ERB frequency scales, *Digital Signal Processing* **104** (2020), 102763, DOI: 10.1016/j.dsp.2020.102763.

[12] M. Swain, A. Routray and P. Kabisatpathy, Databases, features and classifiers for speech emotion recognition: a review, *International Journal of Speech Technology* **21** (2018), 93 – 120, DOI: 10.1007/s10772-018-9491-z.