Fusing Fixed and Adaptive Multi-resolution Features: A DWT-EWT Approach for Improved Speech Emotion Classification

Devi Prasad Pattnaik* and Bala Sai Srilatha Indira Dutt Vemuri

Department of EECE, GITAM Deemed to be University, Visakhapatnam
dpattnai@gitam.edu*, svemuri@gitam.edu

Abstract. Speech emotion classification (SEC) is the automatic identification process of the emotional states that are inherent parts of any utterance with the help of computer programming with high potential applications in the domain of medicine, security, surveillance, digital marketing, E-learning, internet search, personal communication, customer relation mechanisms, human-computer interaction, etc. Recent advances in speech emotion classification performance (ECP) employed various acoustic as well as non-acoustic features with the help of machine learning as well as deep learning algorithms. This paper introduces a new computing mechanism with the help of the hybrid approach of Discrete Wavelet Transform (DWT) and Empirical Wavelet Transform (EWT) with the intention to increase the classification accuracy level with the help of the proposed hybrid signal decomposition and feature fusion technique. Speech signals are broken into frames, that are then decomposed into four modes with the help of the proposed approach. i.e using DWT and EWT, followed by the extraction of five different entropy-based features namely "Approximate Entropy (ApE)", "Permutation Entropy (PrE)", "Increment Entropy (InE)", "Sample Entropy (SaE)", "Spectral Entropy(SpE)", collectively termed Hybrid-Entropy (HEn) features and HMFCC features(Hybrid-Mel-Frequency Cepstral Coefficient). Experimental evaluation using a deep neural network (DNN) classifier demonstrates that combining HEn with HMFCC features derived from both decomposed modes achieves superior performance, attaining an accuracy of 89.76% on the EMODB dataset.

Keywords: DWT, EWT, SEC, Feature Fusion, Multi-resolution Analysis

1 Introduction

It is observed that speech emotion classification has risen significantly due to the advantages of utilizing speech as an emotional biomarker, as mentioned in the research article [7]. Hardware requirements do not affect the usage of speech as an emotional signal, as it is easily accessible and has consistent emotional signals regardless of the language being spoken. SEC is commonly used in various applications like adaptive education, affective care, customer service, etc. [6].

S. Sharma et al. (eds.), *Proceedings of the International Conference on Recent Advances in Intelligent and Sustainable Technologies (RAIST 2026)*, Atlantis Highlights in Intelligent Systems 17,
https://doi.org/10.2991/978-94-6239-707-1_18

The success of an SEC system depends on two major components: the feature set that provides a representation of emotions and a classifier that maps these representations to emotion states [3]. Although most of these methods in classical emotion recognition paradigms involve using handcrafted acoustic features like MFCC [5] and classifiers such as SVM [4], in more recent research on emotion recognition using deep machines has shown superiority in efficiently modeling non-linear dynamics of emotion in speech signals [2]. The researcher in this study has used a DNN classifier because of its compromising characteristics between representational power and efficiency in emotion recognition tasks.

1.1 Motivation

The necessity of combining DWT and EWT in speech emotion classification is grounded in their relative advantages in describing a complex signal that varies over time and exhibits a rich variety of emotion-related content from a different point of view. Although DWT can provide a consistent multi-resolution analysis of a signal in describing emotion-related content of different detail levels in pre-specified frequency ranges, DWT basis functions might not function in an optimum way in describing speaker- and emotion-dependent characteristics of a complex speech signal in different frequency ranges. EWT bases its filter bank on a given signal's spectrum in Fourier representation and has shown potential in extracting emotion-related content in frequency ranges that can vary from utterance to utterance [22]. Therefore, there is a potential in combining both transforms in search of a more robust and discriminative set of features that can reasonably classify emotion-related content in a speech signal [9, 11].

1.2 Key Contributions

The key contributions of our works are as follows:

- Preprocess the speech signal and divide itinto number of frames(50 ms duration with 50% overlap)
- Decompose every speech frame into 4 sub-bands using DWT and 4 modes using EWT.
- Extract 13-MFCC features from every decomposed modes. (Total $13 \times 8 = 104$ HMFCC features per speech frame)
- Five different types of entropy-based features ApE(2), PrE(2), InE(1), SaE(1), and SpE(2) are derived from each decomposed modes, collectively termed HEn(10) features. (Total $10 \times 8 = 80$ HEn features speech frame)
- Concatenate all extracted HMFCC and HEn features extracted from each and every mode.
- Use of DNN classifier to test, train and classify emotions and evaluation of speech emotion classification performance (ECP) i.e classification accuracy estimation.

2 Related Works

Several studies have explored diverse acoustic feature sets for speech emotion recognition. For instance, Aouani et.al [25] employed conventional features such as MFCC, PLP, and TEO, while [15] focused on modulation spectral features. [14] combined Fourier parameters with MFCC, and [13] proposed Gabor spectrogram based local Hu moments. Similarly, [12] utilized residual sinusoidal peak amplitude (RSPA) along with MFCC. Wavelet packet–based features were explored by [11] and [9], whereas [10] employed a richer set, including pitch, intensity, percentiles, formants, MFCC and their derivatives, as well as filter bank energies. Finally, [8] extracted MFCC in combination with ZCR, pitch, and energy-based features. But their analysis considered the complete speech waveform, disregarding separate frequency band emphasis. A study done by Mishra et. al [26] enhances emotion classification performance (PECL) by integrating variational mode decomposition and Hilbert transform approaches to extract unique HT-based entropy features with MFCCs. Experimental results show that combining MFCC and hilbert transform entropy features with a DNN classifier improves SER accuracy to 86.92% for the EMOVO dataset.

3 Dataset Used

The Berlin Emotional Speech Database (EMO-DB) is one of the most used benchmark datasets for speech emotion recognition research. It was made by the Technical University of Berlin and has recordings of ten professional German actors (five male and five female). The dataset includes around 535 utterances expressed in seven different emotional categories: anger, boredom, disgust, fear, happiness, sorrow, and neutrality. To assure emotional authenticity, all recordings are studio-quality, made in a controlled atmosphere, and thoroughly evaluated using perceptual listening tests. Because of its balanced design, high-quality recordings, and well-annotated labels, EMO-DB has become a common reference corpus for testing and comparing speech emotion identification techniques.

4 Methodology

4.1 Pre Processing

It consists of basically three steps in our work. They are preemphasis, normalization and segmentation of speech signals into small frames of brief duration. Each divided frames are of 50 ms with 25 ms overlap.

4.2 Discrete Wavelet Transform(DWT)

The DWT approach decomposes speech signals into sub-bands by passing them through low pass filter (LPF) and High pass filter (HPF). Let $I_L(n)$, $O_L(n)$

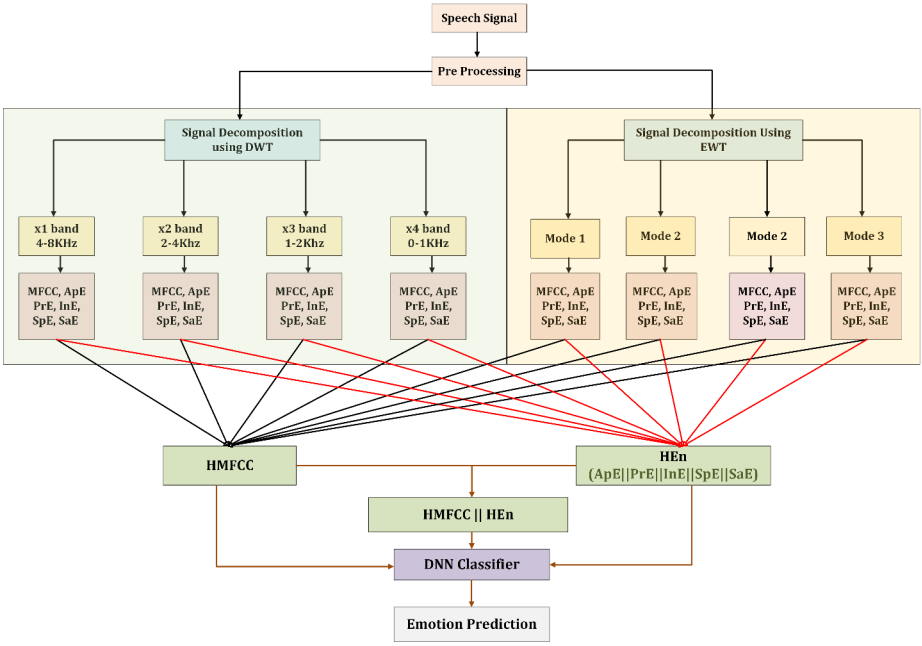

Fig. 1. Block Diagram of the proposed methodology

impulse response and output of LPF and $I_H(n)$, $O_H(n)$ represent the impulse response and output of the HPF.

$$O_L(n) = x(n) * I_L(n) = \sum_{m=-\infty}^{\infty} x(m) I_L(n-m) \tag{1}$$

$$O_H(n) = x(n) * I_H(n) = \sum_{m=-\infty}^{\infty} x(m) I_H(n-m) \tag{2}$$

where $x(n)$ is speech signal frame. Our work uses Daubechies 4 (db4) as a basis function. The DWT equations 3 and 4 [1] are used to calculate the LPF and HPF filter coefficients $I_{L,m}$and $I_{H,m}$

$$\xi(n) = \sqrt{2} \sum_{m} I_{L,m} \xi(2n-m) \tag{3}$$

$$\chi(n) = \sqrt{2} \sum_{m} I_{H,m} \chi(2n-m) \tag{4}$$

$$S(n) = \sum_{m} A_{d_0,m} \xi_{d_0,m}(n) + \sum_{d=1}^{d=d_0} \sum_{m} B_{d,m} \chi_{d,m}(n) \tag{5}$$

$\sqrt{2}\xi(2n-m)$ and $\chi(n)$ are composites of the orthonormal basis functions that correspond to the scaling function and wavelet, respectively [1].

4.3 Empirical Wavelet Transform (EWT)

The Empirical Wavelet Transform (EWT) is an adaptive signal decomposition methodology designed to address the limitations of standard wavelet transforms. Traditional wavelet approaches use fixed basis functions, which may not capture the spectrum properties of highly dynamic sources like speech. On one hand, in the case of EWT, the Fourier spectrum of the input data is divided into a number of consecutive bands in accordance with their content. Then, an empirical wavelet filter bank is constructed for these bands, ensuring the successful decomposition of the data into effective modes. With such an empirical method, it is possible to achieve better localization for the data in the time-frequency plane; hence, the method is suitable for the evaluation of non-linear and non-stationary data, e.g., for the task of speech emotion identification.

The first step of EWT involves partitioning the Fourier spectrum of the signal into a set of contiguous frequency bands $\{\Lambda_n\}_{n=0}^{N}$ based on significant transitions in the spectrum. For each frequency band $\Lambda_n = [\omega_n, \omega_{n+1}]$, an empirical scaling function $\hat{\phi}(\omega)$ and an empirical wavelet function $\hat{\psi}_n(\omega)$ are constructed as follows:

$$\hat{\phi}(\omega) = \begin{cases} 1, & |\omega| \leq \omega_0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$$\hat{\psi}_n(\omega) = \begin{cases} 1, & \omega_n \leq |\omega| \leq \omega_{n+1} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

The signal $x(t)$ can then be decomposed into empirical modes by filtering with these empirical wavelets:

$$x(t) = \sum_{n=0}^{N} x_n(t), \quad x_n(t) = \mathcal{F}^{-1}\left(\hat{\psi}_n(\omega) \cdot \hat{x}(\omega)\right) \tag{8}$$

$\hat{x}(\omega)$ is the Fourier transform of $x(t)$, and $\mathcal{F}^{-1}$ is the inverse Fourier transform. Each $x_n(t)$ represents an empirical mode associated with a separate spectral band. Each EWT mode is is represented by $x_n(t)$.

4.4 Feature Extraction

The effectiveness of SEC is strongly influenced by the choice of representative features. Accurate emotion discrimination requires selecting features that convey the maximum amount of emotional information. In our work 104 HMFCC features and 80 HEn features are extracted from the decomposed modes that is later concatenated and fed to DNN. A total of 184 features are extracted from a speech signal frame.

Mel-Frequency Cepstral Coefficients (MFCC) represent the short-term spectral envelope of speech on the perceptual Mel scale and are obtained by applying

the Discrete Cosine Transform (DCT) to the log-Mel filterbank energies:

$$c_n = \sum_{m=1}^{M} \log(S_m) \cdot \cos\left(\frac{\pi n}{M}(m - 0.5)\right), \quad n = 1, 2, \ldots, K, \tag{9}$$

where S_m denotes the Mel-scaled filterbank energy. The Mel-Spectrogram provides a time–frequency representation in the Mel domain and is expressed as

$$\text{MEL}(t, m) = \sum_{k=0}^{N-1} |X_t(k)|^2 \cdot H_m(k), \tag{10}$$

with $X_t(k)$ being the STFT of frame t and $H_m(k)$ the Mel filter response. AE quantifies signal regularity based on the likelihood that similar patterns of length m remain similar for length $m + 1$, defined as

$$ApE(m, r, N) = \phi^m(r) - \phi^{m+1}(r). \tag{11}$$

PE evaluates the complexity of a time series using the distribution of ordinal patterns, computed as

$$PrE(m) = -\sum_{i=1}^{m!} p(\pi_i)\, \log\big(p(\pi_i)\big), \tag{12}$$

where $p(\pi_i)$ is the probability of ordinal pattern π_i.

IE evaluates complexity based on the distribution of symbolic patterns derived from incremental changes .

$$InE = -\sum_{j=1}^{S} p(s_j)\, \log_2\big(p(s_j)\big) \tag{13}$$

where S is the total number of distinct increment patterns $p(s_j)$ is the probability of increment pattern s_j.

SpE measures the flatness of the power spectrum and is calculated as

$$SpE = -\sum_{i=1}^{N} P_i\, \log_2(P_i) \tag{14}$$

with $P_i = \frac{|X(f_i)|^2}{\sum_j |X(f_j)|^2}$ representing the normalized spectral power at frequency f_i.

Similarly SE measures the irregularity of a time series as the negative log probability that similar sequences of length m remain similar at length $m + 1$.

$$SaE(m, r, N) = -\ln\left(\frac{A}{B}\right) \tag{15}$$

where A and B are the counts of matched sequences of length $m + 1$ and m within tolerance r, respectively.

4.5 DNN Classifier

The illustrated Deep Neural Network (DNN) architecture as shown in table 1.The model includes four hidden layers configured with 1024 neurons in the first layer, 512 neurons in the second and third layers, and 256 neurons in the fourth layer. To reduce overfitting, the hidden layer applies a dropout rate of 0.1 and 0.2. Activation functions are carefully selected: ELU (Exponential Linear Unit) is used in the first, second, third, and fourth layers. Finally, the output layer contains seven neurons with a SoftMax activation function, which maps the learned features to seven emotion classes, producing the predicted emotion with high accuracy.

The DNN architecture used and the total trainable parameter is shown in table 1. Out of 535 samples 80% used for training and 20% used for testing.

Table 1. Proposed DNN architecture and parameter count using an enhanced DWT-EWT feature set.

Layer	**Type**	**Neurons**	**Dropout**	**Activation**	**Trainable Parameters**
Input	Feature Vector	184	–	–	–
1	Hidden	1024	0.1	ELU	189,440
2	Hidden	512	0.1	ELU	524,800
3	Hidden	512	0.2	ELU	262,656
4	Hidden	256	0.2	ELU	131,328
5	Output	7	–	SoftMax	1,799
Total Trainable Parameters					**1,110,023**

5 Results and Discussion

In this part, we used the suggested features HMFCC and HEn to compare the ECP for the EMODB dataset. The HMFCC and HEn characteristics were used both singly and together as input features for a DNN classifier in our study to evaluate the ECP. For up to 800 epochs with the EMODB dataset, it was seen that there were no improvement in classification accuracy beyond 750 epochs. Table 2 presents the evaluation outcomes for the proposed features assessed individually and together with a DNN classifier on the EMODB dataset. The experiments reveal that the fusion of HEn and HMFCC features leads to improvements in classification accuracy (ECP).

The confusion matrix heatmap shown in figure 2 depicts the efficacy of emotion categorization on the EMODB dataset. Several emotions, including anger, neutral, and sadness, demonstrate high recognition rates of more than 90%. Moderate misclassifications exist for happiness and, to a lesser extent, fear, with these categories overlapping more with boredom and anger. Most emotion classes have

Table 2. Performance of the proposed feature for the EMODB speech dataset

Proposed Feature	Proposed Feature Dimension	Classification Accuracy (%)
HMFCC	104	85.35
HEn	80	80.26
MFCC + HEn	184	89.76

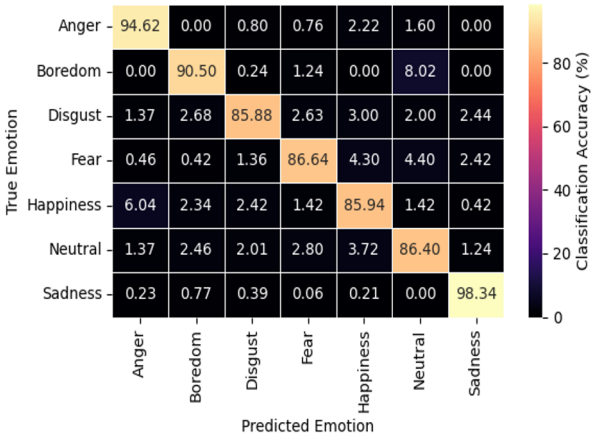

Fig. 2. Confusion Matrix Heatmap (HMFCC + HEn) – EMODB

low off-diagonal errors, indicating strong feature discrimination and successful separation of diverse emotions. However, happiness is easily mistaken with other emotional states, implying that this class need additional refining. Overall, the heatmap demonstrates the classifier's high accuracy for most emotions while also identifying areas for improvement in recognition where categories overlap.

6 Comparison with Similar Methods

As shown in Table 3, earlier approaches on the EMODB dataset achieved accuracies ranging from 72.38% [16] to 84.86% [23]. In comparison, the proposed method combining spectral (HMFCC) and entropy-based (HEn) features taken from both DWT and EWT decomposed modes with a DNN attained the highest accuracy of 89.76%, outperforming all existing methods.

7 Conclusion

Automated emotion recognition from speech cues is a complex and difficult task. Researchers have investigated a variety of ways to improve classification performance, including feature or classifier-level fusion, the integration of distinct acoustic and non-acoustic information, and the use of several machine or deep

Table 3. Performance comparison of the proposed method on the EMODB dataset

Authors	**Features**	**Classifier**	**Acc. (%)**
Latif et al. [16]	eGeMAPS	DBN	72.38
Pham et al. [17]	Spectral combo	1D-CNN	76.40
Assunção et al. [18]	VGGVox	LMT	80.40
Ancilin et al. [19]	MFMC	SVM	81.50
Pandey et al. [20]	MFCC	CNN+BLSTM	82.35
Flower et al. [21]	RFT	SVM	83.08
Pattnaik et al. [24]	MRD-MFCC, MRD-Mel, MRD-PE, MRD-AE	DNN	84.01
Mishra et al. [23]	$MFCC_{mean}$, $MFCC_{SE}$, 13 coefficient	DNN	84.86
Proposed	**HMFCC + HEn**	**DNN**	**89.76**

learning models. In our research, we present a novel technique that combines DWT and EWT to improve ECP. Our strategy of combining HMFCC features with HEn features and using a DNN classifier achieved a highest accuracy of 89.98%. This enhancement demonstrates the efficacy of combining fixed and adaptive multi resolution based features to increase the discriminative capabilities of emotion classification models.

References

1. Vetterli, M.: Wavelets and Subband Coding. Prentice Hall (1995)
2. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., Zafeiriou, S.: Adieu Features? End-to-End Speech Emotion Recognition Using a Deep Convolutional Recurrent Network. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200–5204. IEEE (2016).
3. Ververidis, D., Kotropoulos, C.: Emotional Speech Recognition: Resources, Features, and Methods. Speech Communication, 48(9), 1162–1181, 2006.
4. Reynolds, D. A., Quatieri, T. F., Dunn, R. B.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10(1–3), 19–41, 2000.
5. Eyben, F., Scherer, S., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S. S., Truong, K. P.: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing 7(2), 190–202, 2016
6. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T. F.: A Review of Depression and Suicide Risk Assessment Using Speech Analysis. Speech Communication 71, 10–49, 2015.
7. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends. Communications of the ACM **61**(5), 90–99, 2018.
8. Avots, E., Sapiński, T., Bachmann, M., Kamińska, D.: Audiovisual Emotion Recognition in Wild. Machine Vision and Applications 30(5), 975–985, 2019.

9. Palo, H. K., Mohanty, M. N.: Wavelet Based Feature Combination for Recognition of Emotions. Ain Shams Engineering Journal 9(4), 1799–1806, 2018.
10. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Audio-Visual Emotion Recognition in Video Clips. IEEE Transactions on Affective Computing **10**(1), 60–75, 2017.
11. Huang, Y., Wu, A., Zhang, G., Li, Y.: Extraction of Adaptive Wavelet Packet Filter-Bank-Based Acoustic Feature for Speech Emotion Recognition. IET Signal Processing 9(4), 341–348, 2015.
12. Deb, S., Dandapat, S.: Emotion Classification Using Residual Sinusoidal Peak Amplitude. In: International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IEEE, 2016.
13. Tao, H., Liang, R., Zha, C., Zhang, X., Zhao, L.: Spectral Features Based on Local Hu Moments of Gabor Spectrograms for Speech Emotion Recognition. IEICE Transactions on Information and Systems 99(8), 2186–2189, 2016.
14. Wang, K., An, N., Li, B. N., Zhang, Y., Li, L.: Speech Emotion Recognition Using Fourier Parameters. IEEE Transactions on Affective Computing 6(1), 69–75, 2015.
15. Sharma, R., Vignolo, L., Schlotthauer, G., Colominas, M. A., Rufiner, H. L., Prasanna, S. R. M.: Empirical Mode Decomposition for Adaptive AM-FM Analysis of Speech: A Review. Speech Communication 88, 39–64, 2017.
16. Latif, S., Rana, R., Younis, S., Qadir, J., Epps, J.: Transfer Learning for Improving Speech Emotion Classification Accuracy. arXiv preprint arXiv:1801.06353, 2018.
17. Pham, M. H., Noori, F. M., Torresen, J.: Emotion Recognition Using Speech Data with Convolutional Neural Network. In: 2021 IEEE 2nd International Conference on Signal, Control and Communication (SCC), pp. 182–187. IEEE 2021.
18. Assunção, G., Menezes, P., Perdigão, F.: Speaker Awareness for Speech Emotion Recognition. International Journal of Online and Biomedical Engineering **16**(4), 15–22, 2020.
19. Ancilin, J., Milton, A.: Improved Speech Emotion Recognition with Mel Frequency Magnitude Coefficient. Applied Acoustics 179, 108046, 2021.
20. Pandey, S. K., Shekhawat, H. S., Prasanna, S. R. M.: Deep Learning Techniques for Speech Emotion Recognition: A Review. In: 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), pp. 1–6. IEEE 2019.
21. Flower, T. M. L., Jaya, T.: Speech Emotion Recognition Using Ramanujan Fourier Transform. Applied Acoustics 201, 109133, 2022.
22. Gilles, J.: Empirical Wavelet Transform. IEEE Transactions on Signal Processing 61(16), 3999–4010, 2013.
23. Mishra, S. P., Warule, P., Deb, S.: Speech Emotion Recognition Using MFCC-Based Entropy Feature. Signal Image and Video Processing 18, 153–161, 2024.
24. Pattnaik, D. P., Vemuri, B. S. S. I. D.: Sub-Band Based Analysis for Speech Emotion Classification. Applied Acoustics 239, 110821 2025.
25. Aouani, Hadhami and Ayed, Yassine Ben, Speech emotion recognition with deep learning, Procedia Computer Science, Elsevier vol 176, pp 251–260, 2020.
26. Siba Prasad Mishra, Pankaj Warule, Suman Deb, Speech emotion recognition using a combination of variational mode decomposition and Hilbert transform, Applied Acoustics, Volume 222, 110046, 2024.